Portuguese Version

A Inteligência Artificial como Fator Positivo e Negativo dentro do Risco Global.

Em breve no Global Catastrophic Risks, eds. Nick Bostrom e Milan Cirkovic

Paper de 31 de Agosto de 2006

Eliezer Yudkowsky

(yudkowsky@singinst.org)

Singularity Institute for Artificial Intelligence

Palo Alto, CA

Intyrodução

O maior perigo da AI (Inteligência Artificial) é, sem dúvida, as pessoas acharem, cedo demais, que já entendem tudo sobre o assunto. Obviamente esse problema não está contido somente nessa campo. “Um curioso aspecto da Teoria da Evolução é que todo mundo acha que a entende”, disse Jacques Monod (Monod 1974). Meu pai, físico, reclamava que as pessoas gostavam de fazer suas próprias teorias de física, ele se perguntava por que elas não faziam de química. (Elas fazem). De qualquer modo o problema parece ocorrer mais comumente no caso da AI. A área da AI tem a má reputação de prometer mundos e fundos e depois não cumprir. A maioria conclui que AI é difícil, como de fato o é, mas o constrangimento não vem da dificuldade em si. Certamente é difícil construir uma estrela, mesmo tendo hidrogênio suficiente, mas a área de astrofísica não tem essa péssima reputação de prometer fazer estrelas e não entregar. A dedução crítica não é que a AI seja difícil, mas que, por algum motivo, seja muito fácil as pessoas acharem que saibam muito mais sobre o assunto do que realmente o façam.

O meu outro capítulo de *Global Catastrophic Risks, "Cognitive biases potentially affecting judgment of global risks",* foi aberto com uma observação que poucas pessoas escolheriam deliberadamente destruir o planeta em que vivemos; mas um cenário em que a destruição ocorre por um erro, ou seja, sem dolo, é realmente preocupante. Poucas pessoas apertariam um botão que elas soubessem claramente que causaria uma catástrofe global, mas se pessoas de boa fé acreditassem que aquele botão faria algo bem diferente, isso sim é causa para alarme.

Entendo ser muito mais difícil escrever sobre AI que sobre Vieses Cognitivos, pois esses últimos são ciências já desenvolvidas e maduras, basta citar a literatura existente. AI não é uma ciência desenvolvida, ela pertence à fronteira, não aos manuais. E, por razões já discutidas em sessões anteriores no tópico “Riscos Catastróficos Globais da AI”, não há, praticamente, discussões na literatura técnica existente. Eu analisei, forçosamente portanto, a matéria a partir de minha própria perspectiva; chegando às minhas conclusões e apresentando meus argumentos, que buscam sustentá-las, da melhor maneira possível dentro do limitado espaço que dispomos. Não que eu tenha sido negligente em citar os trabalhos relevantes nesse tópico. A questão é que, no meu melhor discernimento, não há trabalhos relevantes a ser citados (isso em janeiro de 2006).

NOTA 1 I thank Michael Roy Ames, Eric Baum, Nick Bostrom, Milan Cirkovic, John K Clark, Emil Gilliam, Ben Goertzel, Robin Hanson, Keith Henson, Bill Hibbard, Olie Lamb, Peter McCluskey, and Michael Wilson for their comments, suggestions and criticisms. Needless to say, any remaining errors in this paper are my own.

É, de certo modo, tentador simplesmente ignorar a AI, pois de todos os riscos globais discutidos neste livro, AI é o de mais difícil debate. Não temos estatísticas atuariais para consultar e poder dimensionar pequenas probabilidades anuais de ocorrência de catástrofe, do modo como é feito com choque de asteróides. Não podemos usar cálculos precisos, ou modelos precisamente confirmados para descartar hipóteses baseados em suas probabilidades como alguns desastres físicos, mas isso faz das catástrofes com AI mais preocupantes e não menos.

O efeito de muitos vieses cognitivos tem efetivamente aumentado com o tempo e informações esparsas. O que significa dizer que quanto mais difícil o desafio analítico, mais importante será evitar ou reduzir os vieses. Portanto eu recomendo fortemente a leitura do "Cognitive biases potentially affecting judgment of global risks", pp. XXX-YYY, antes de continuar com esse capítulo.

1: Viés Antropomórfico

Quando algo é muito comum em nosso dia a dia, nós o damos como certo a tal ponto de esquecer-se de sua existência.

Imagine uma adaptação biológica complexa que necessite dez partes. Se cada um dos dez genes tem, independentemente, uma freqüência de 50% no pool de genes – cada gene aparece somente em metade dos indivíduos daquela espécie – então, na média, somente 1 em cada 1024 individuos terá essa adaptação completamente funcional. Um revestimento de pelos não é uma vantagem evolucionária muito relevante se o meio não desafiar, seguramente, os indivíduos com o frio. Similarmente, se o gene B depende do gene A, então o gene B não tem vantagem significativa se o gene A não estiver presente de forma regular no ambiente genético. Um mecanismo interdependente e complexo é necessariamente universal para espécies que se reproduzem de forma sexuada, de outra maneira, não poderia evoluir. (Tooby and Cosmides 1992) Um pardal pode apresentar penas mais brilhantes que outro, mas ambos terão asas. A seleção natural, enquanto se alimenta das variações, a exauri, (Sober 1984).

Em toda cultura, os humanos experimentam alegria, tristeza, desgosto, ira, medo e surpresa (Brown 1991), e demonstram essas emoções através das mesmas expressões faciais (Ekman e Keltner 1997). Todos rodamos o mesmo motor sob o capô, mas gostamos que sejam pintados com cores diferentes; um princípio que os psicólogos evolucionários chamam de unidade psíquica da humanidade (Tooby e Cosmides 1992). Essa observação é duplamente explicada e requerida pelos mecanismos da biologia evolucionária.

Um antropologista não ficará muito excitado em anunciar uma descoberta de uma nova tribo: “Eles comem alimentos! Eles respiram ar! Eles usam ferramentas! Eles contam casos entre si!” Nós humanos nos esquecemos como somos parecidos uns com outros, vivendo num mundo que somente nos lembra de nossas diferenças.

Os humanos evoluíram copiando outros humanos, competindo e cooperando com seus semelhantes. Era de se esperar que nesse ambiente ancestral qualquer inteligência encontrada estivesse somente em outro humano. Nós evoluímos compreendendo nossos companheiros com empatia, colocando-nos no lugar do próximo; pois aquilo que precisava ser copiado ou moldado era similar ao copiador ou moldador. É, portanto, natural que os seres humanos tendem, na maioria das vezes, a uma “antromorfização”, ou seja, atribuir características humanas àquilo que não é humano. No filme Matrix (Wachowski e Wachowski 1999), a suposta

“inteligência artificial” agente Smith se apresenta bastante relaxado, passivo e sem emoção. Mas depois, enquanto interroga o humano Morfeu, o agente deixa transparecer todo seu desgosto com a humanidade, e sua face mostra a universalmente humana face do desgosto.

Indagar seu próprio cérebro pode funcionar bem como uma adaptação instintiva, se a idéia for prever o comportamento de outros humanos. Mas, se você estiver tratando com qualquer outro tipo de processo de otimização, como, por exemplo, se você fosse o teólogo do sec. XVIII William Paley, olhando para a complexa ordem da vida e se perguntando como tudo isso apareceu, então o antropomorfismo é uma armadilha tão perigosa para os cientistas incautos cairem facilmente, que só um Darwin poderia escapar.

Experiências sobre o antropomorfismo mostram que as cobaias o fazem inconscientemente, na maioria das vezes voando na cara de suas deliberadas crenças. Num estudo realizado por Barret e Keil (1996), as cobaias mostraram fortes crenças nas propriedades não-antropomórficas de Deus: que Deus poderia estar em mais de um lugar ao mesmo tempo, ou prestar atenção a múltiplos eventos simultanemamente. Barrett e Keil apresentaram às cobaias estórias em que, por exemplo, Deus salvou pessoas de afogamento. As cobaias responderam questões sobre essas estórias e então as recontaram em suas próprias palavras de forma que Deus estava em apenas um lugar a cada momento e executou as tarefas sequencialmente e não simultaneamente. Felizmente para nossos objetivos, Barrett e Keil também testaram outro grupo usando de outra forma as mesmas estórias sobre um computador superinteligente chamado “Umcomp”. Por exemplo, para simular a propriedade de onipresença as cobaias foram informadas que os sensores e medidores “cobriam cada centímetro quadrado da Terra, portanto nenhuma informação escaparia do processamento”. Esse grupo também mostrou forte antropomorfismo, mas significativamente menos que o primeiro grupo (o de Deus). Da nossa perspectiva, o resultado chave é que mesmo quando as pessoas conscientemente crêem que uma AI não seja como um humano, elas ainda assim visualizam cenários como se a AI fosse antropomórfico (mas não tão antropomórfico quanto Deus).

O viés antropomórfico pode ser classificado como insidioso: ele aparece sem intenção deliberada, sem percepção consciente, mas implícito no conhecimento aparente.

Na época de ficção científica de nível B, as capas dos livretos mostravam ocasinalmente um monstruoso, mas perceptivo alienígena, coloquialmente conhecido por monstro do olho de inseto ou MOI, carregando uma mulher

atraente vestida em saias rasgadas. Aparentemente o desenhista acreditava que um alienígena não-humanóide, com uma história evolucionária completamente diferente da nossa, teria um desejo sexual por fêmeas humanas. As pessoas não fazem erros desse tipo explicitamente raciocinando como: “Todos os tipos de mentes funcionam mais ou menos da mesma forma, portanto, presumivelmente, um MOI deve achar nossas mulheres atraentes.” Provavelmente o desenhista nem se perguntou se um inseto gigante acha as nossas mulheres atraentes. Mas sim pensou: ”Uma mulher em saia rasgada é sexy”, inerentemente, tanto quanto uma proriedade intrínseca. Os desenhistas daqueles tempos que cometeram esses erros não pensaram na mente do superinseto, eles focaram na saia rasgada da mulher. Se a saia não estivesse rasgada a mulher seria menos sexy. O MOI não entra nessa discussão. (Esse é um caso de equívoco grave, confuso e extraodinariamente comum que E. T. Jaynes o nomeou de falácia da projeção mental. (Jaynes e Bretthorst, 2003.) Jaynes um teórico da probabilidade Bayesiana, usou “falácia da projeção mental” para se referir a erros de estado de conhecimento com propriedades dos objetos. Por exemplo, o termo “fenômeno misterioso” implica que o mistério seja uma propriedade do fenômeno. Se eu sou ignorante sobre o fenômeno, então isso é um fato relativo ao meu estado de consciência, não ao fenômeno em si.)

As pessoas não precisam perceber que estão antropomorfizando (ou mesmo se dar conta que estão excutando um questionável ato de prever o comportamento de outras mentes) para que o antropomorfismo se sobreponha à cognição. Quando tratamos de refletir sobre outras mentes, cada passo no processo de raciocínio pode ser contaminado por suposições tão ordinárias na experiência humana que nos passa tão despercebidas quanto o ar que respiramos ou a gravidade. Você faz a seguinte objeção ao ilustrador do livreto: ”Não seria mais provável que um inseto gigante macho tivesse atração sexual por um inseto gigante fêmea?” O ilustardor pensa por um instante e diz então a você: “Bem, se um alienígena em forma de inseto normalmente gosta de exoesqueletos duros, ele pode mudar de idéia, depois de conhecer uma fêmea humana e perceber que ela tem uma pele suave e admirável, e então, se for suficientemente avançado em sua tecnologia, pode se reprogramar geneticamente para começar a gostar de pele suave em vez de exoesqueleto duro.”

Esta é uma Falácia tipo “toque de retirada”. Depois que o pensamento antropomórfico do alien é apresentado, o ilustrador dá um passo para trás e tentar justificar sua conclusão como um produto neutro do processo mental alienígena. Talvez o avançado MOI possa, de fato, fazer uma reengenharia de si mesmo (geneticamente ou de outra forma) para gostar de pele suave, but será que ele deseja fazê-lo? Um MOi que gosta de exoesqueletos duro não deseja introduzir mutações em si mesmo para começar a gostar de pele suave, a não ser que a seleção natural produzido nele um distinto senso de metassexualidade humana. Quando usamos uma cadeia longa de complexa de raciocínio para justificar conclusões antropomórficas, cada passo do raciocínio é uma oportunidade para cair no erro.

Também vem a ser um sério erro começar a partir da conclusão e buscar uma linha de raciocínio aparentemente neutra que o levará até lá; isso é racionalização. Se for uma autossugestão a responsável aquela imediata imagem mental de um inseto gigante perseguindo uma mulher, então o antropomorfismo é a principal causa dessa crença, e nenhuma quantidade de racionalização mudará isso.

Qualquer pessoa buscando reduzir o viés antropomórfico em si mesma deve ser alertada para estudar biologia evolutiva a fim de praticar, preferentemente biologia evolutiva com matemática. Os primeiros biologistas, muitas vezes antropomorfizavam a seleção natural, eles acreditavam que a evolução faria as mesmas coisas que eles fariam; eles tentaram prever os efeitos da evolução colocando-se no lugar da evolução. O resultado foi uma enorme quantidade de conclusões sem o menor sentido que começaram a ser retiradas do campo da biologia no final da década de 60, e.g. por Williams (1966). A biologia evolutiva oferece tanto matemática como estudos de *cases* que ajudam a extrair o viés antropomórfico.

1.1: The width of mind design space

A evolução conserva algumas estruturas de forma extraordinária. Uma vez que um gene evolue dependendo de outro gene já existente, esse outro gene está amarrado; ele não pode sofrer mutações sem quebrar uma série de adaptações. Genes homeóticos, aqueles que controlam o desenvolvimento de um corpo nos embriões, determinam quando muitos outros genes devam se manisfestar (ativar). A mutação num gene homeótico pode resultar em um embrião da mosca da fruta se desenvolver normalmente, exceto que sem cabeça. Como resultado os genes homeoticos são guardados a sete chaves pela evolução natural a tal ponto que muitos deles são os mesmos em moscas e humanos, eles não sofreram mutações desde o último ancestral comum entre as moscas e o homo sapiens. O mecanismo molecular empregado na síntese do ATP é essencialmente o mesmo, seja em mitocôndrias de animais, em cloroplastos de plantas ou em bactérias; A síntese do ATP não sofreu mutações significativas desde o aparecimento das céluas eucarióticas há dois bilhões de anos passados.

Dois designs quaisquer de AIs podem ser menos similares entre si que você o é a uma petúnia.

O termo “Inteligência Artificial” se refere a uma gama de possibilidades muito mais abrangente que o termo ”Homo Sapiens”. Quando falamos sobre Ais estamos, na verdade, falando de mentes em geral, ou processos de otimização em geral. Imagine um enorme mapa do espaço de design de mentes. Num canto um pequeno círculo que contenha todas as mentes

humanas, que por sua vez esteja dentro de outro circulo um pouco maior que contenha toda a vida biológica; todo o resto do imenso mapa é o espaço das mentes em geral. Esse mapa flutua inteiramente num espaço ainda maior, o espaço dos processos de otimização. A seleção natural cria mecanismos funcionais complexos sem a ajuda da mente; a evolução reside no espaço dos processos de otimização, mas fora do círculo das mentes em geral.

É esse enorme espaço de possibilidades que exclui o antropomorfismo como um raciocínio legitimado.

2: Previsões e design

Nãopodemos indagar nossos próprios cérebros sobre questões de processos de otimização nãohumanos, seja MOI, seleção natural ou AIs. Como devemos proceder então? Como podemos prever o que as AIs farão? Eu, deliberadamente fiz essa questão de uma forma que a torna intratável. Parando o problema, é impossível prever se um sistema computacional com liberdade para decidir irá implementar alguma função de entrada-saída de dados, incluindo digamos, uma simples multiplicação (Rice 1953). Então como é possível que engenheiros humanos sejam capazes de construir chips de computadores que com certeza irão implementar multiplicação? É por que os engenheiros humanos usam deliberadamente designs que eles possam compreender.

Antropomorfismo leva as pessoas a acreditarem que elas podem fazer previsões, dada nehuma outra informação além de que algo é uma “inteligência”, antopomorfismo seguirá fazendo previsões de qualquer forma, seu cérebro automaticamente se colocando no lugar da “inteligência”. Este pode ser um dos fatores que contribuíram para a constrangedora história das AIs, que nasce não da dificuldade das AIs como tal, mas a partir da misteriosa facilidade de adquirir crenças equivocadas sobre o que um dado design de AI poderá realizar.

Para afirmar com segurança que uma ponte suportará veículos de até 30 ton, os engenheiros civis possuem duas ferramentas: Escolha das condições inciais e margem de segurança. Tudo que precisam fazer é projetar uma única ponte sobre à qual pode-se fazer tal afirmação. E, apesar disso refletir na capacidade de um engenheiro para calcular corretamente peso exato que uma ponte irá suportar, é também aceitável que se calcule que a ponte suporte pelo menos 30 ton, apesar de que para aceitar esta vaga afirmação de forma rigorosa irá requerer muito da compreensão teórica envolvida no cálculo exato.

A engenheiria civil é muito bem conceituada pela sua capacidade de prever que as pontes suportarão veículos. Os alquimistas da antiguidade eram

bem menos conceituados em prever qual sequencia de reações químicas transformaria chumbo em ouro. Quando de chumbo em quanto de ouro? Qual o mecanismo causal? É compreensível a razão pela qual o alquimista prefere ouro em vez de chumbo, mas por que essa tal sequencia transformaria chumbo em ouro, em vez de ouro em chumbo ou chumbo em água?

Os pesquisadores inicialmente acreditavam que uma rede neural artificial em camadas, treinada via propagação reversa seria ‘’inteligente’’. A esperança era, provavelmente, resultante mais da alquemia que da engenharia civil. A ‘’magica’’ está na lista das caractrísticas uneversais feita por Donald Brown (Brown 1991), mas a Ciencia não está. Nós não distinguimos instintivamente entre compreensão rigorosa e um caso bem contado. Nos não notamos institivamente uma expectativa de resultados positivos apoiada no ar.

A espécie humana resulta da seleção natural, que opera através da retenção naoaleatoria das mutações aleatórias. Um caminho que leva a uma catástrofe global, alguém apertando o botão com a idéia equivocada sobre qual sua função, por exemplo, é que a AI vem similarmente através da adição de algoritmos funcionais, sem um profundo entendimento, por parte dos pesquisadores, de como essa combinação irá funcionar. Apesar de não terem uma nítida visualização do exato processo envolvido na geração de um comportamento amigável ou qualquer compreensão detalhada do que eles querem dizer por ‘’ser amigavel’’, eles acreditam que a AI assim o será. Da mesma forma que os pesquisadores iniciais tinham expectativas vagas e equivocadas sobre a ‘’inteligencia’’ de seus programas, nós também imaginamos que eles tiveram sucesso em desenvolver um programa inteligente, mas não temos expectativas vagas e equivocadas que esses programas são amigáveis.

Não saber como construir uma AI amigável não é fatal *per se*, se você sabe que não sabe. A crença equivocada que uma AI será amigável é um obvio caminho para catástrofe global.

3. Subsetimando a força da inteligência

Nós tendemos a ver as diferenças individuais mais que as características universais. Assim quando alguém diz a palavra ‘’inteligencia’’, nós pensamos em Einstein em vez de humanidade.

As diferencas individuais de inteligência humana tem uma medida padrão, o ‘’g’’ de Spearman, também conhecido por ‘’fator-g’’, uma interpretação controversa do solido resultado experimental que diferentes testes de inteligência estao altamente correlacionados entre si e com os eventos da vida real como a renda média durante a vida. (Jensen 1999). O ‘’g’’ de Spearman é uma abstração estatística das diferenças indiduais de inteligência entre humanos, que como uma espécie é muito mais inteligente

que lagartos. O ‘’g’’ de Spearman é abstrato de diferenças milimétricas entre uma espécie de gigantes.

Não devemos confundir o ‘’g’’ de Spearman com a inteligencia humana em geral, nossa capacidade de cumprir uma extensa lista de tarefas cognitivas que são incompreensíveis a outras espécies. A inteligência geral é uma diferença do tipo entre-espécies, uma adaptação complexa e uma característica humana universal encontrada em todas as culturas conhecidas. Pode ainda não haver consenso sobre o que é inteligência, mas não há duvida sobre sua existência ou sua força. Existe algo nos humanos que os permitiram deixar pegadas no solo lunar.

Porém o termo ‘’inteligencia’’ normalmente remete a imagens de um acadêmico (pobre) com um QI de 160 e de um CEO com um QI de meros 120 pontos. De fato há diferenças entre habilidades individuais, além dos pobres acadêmicos, que contribui para o sucesso relativo no mundo humano: entusiasmo, habilidade social, educação, talento musical e racionalidade, por exemplo. Note que os fatores listados são todos cognitivos. Habilidade social reside no cérebro, não no fígado. E, piadas a parte, você não encontrará muitos chimpazés que são CEOs ou professores acadêmicos. Voce não encontrará muitos ratos que são racionalistas, nem artistas, nem poetas e nem músicos compositores. A inteligencia é a fundação do poder humano, a força que move nossas artes.

O perigo de confundir inteligência geral com fator ‘’g’’ é que isso leva a uma tremenda subestimação do potencial impacto da Inteligencia Artificial. (Isso se aplica tanto a impactos socioambientais positivos quanto negativos). Mesmo o termo AI ‘’transhumano’’ ou ‘’superinteligencia artificial’’ pode ainda remeter a imagens de book-smarts-in-a-box: uma AI que é muito eficiente em tarefas cognitivas estereotipicamente associadas com ‘’Inteligencia’’, como xadrez ou matemática abstrata. Mas não superhumanamente persuasiva; ou muito melhor que os humanos em prevê e manipular situações sociais humanoas; ou ainda inhumanamente esperta em formular estratégias de longo prazo. Portanto, em vez de Einstein, deveríamos pensar em, digamos, o gênio da diplomacia e da política do sec. XIX Otto Von Bismark? Mas essa é apenas a visão do erro (invertida) no espelho. O intervalo entre o idiota da vila para Einstein ou do idiota da vila para Bismark é como um ponto em relação o intervalo que separa a ameba do ser humano.

Se a palavra ‘’inteligência’’ evocasse Einstein em vez de humanos, então seria sensível dizer inteligência não pode contra o revolver, como se revolver desse em arvores. Igualmente seria sensível dizer que inteligência não pode contra o dinheiro, como se os ratos utilizassem dinheiro. Os humanos não começaram sua evolução com ativos como garras , dentes, couraças ou qualquer outro tipo de vantagens que eram as moedas de troca de outras espécies. Se olhassemos aos humanos a partir da perspeciva do resto da ecoesfera, ninguém diria que aquelas coisas moles e cor de rosa

iriam, eventualmente, se colocarem dentro de tanques blindados. Nós inventamos o campo de batalha onde vencemos os leões e lobos. Nós não ganhamos o jogo com nossas garras e dentes, nós tínhamos nossas próprias idéias sobre o que realmente importava. Tal é o poder da criatividade.

Vinge (1993) apropriadamente observou que um futuro que tenha mentes mais espertas que as humanas será de diferente espécie. Inteligência Artificial não é um dispositivo extraordinário e caro para ser anunciado em revistas especializadas. Inteligência Artificial não pertence ao mesmo gráfico que mostra o progresso na medicina, manufatura ou energia. Inteligência Artificial não é algo que você pode inserir num cenário pseudofuturista de arranhacéus, automóveis voadores e glóbulos vermelhos nanotecnológicos que lhe permitem segurar a respiração por oito horas seguidas. Arranhacéus suficientemente altos não tem potencial para começar a fazer sua própria engenharia. A humanidade não ascendeu como espécie dominante por que segurava sua respiração por mais tempo que outras espécies.

O cenario catastrófico que vem da subestimação do poder da inteligencia é aquele onde alguém constroi um botão e não se preocupa com o que o botão faz, pois considera-o sem força suficiente para fazer-lhe mal. Ou, a subestimação do poder da inteligência implica proporcionalmente numa subestimação do potencial impacto da Inteligência Artificial. O (ainda pequeno) grupo de pesquisadores, fundraisers e filantropistas envolvidos que manejam riscos existenciais em nome da humanidade não prestarão a devida atenção à Inteligência Artificial. Ou, o amplo campo da AI não prestará a devida atenção aos riscos de uma poderosa AI, portanto não teremos boas ferramentas nem bases firmes para garantir o caráter amigável quando for possível construí-la.

Deve ser mencionado também, pois impacta igualmente no risco existencial, que a Inteligência Artificial poderia ser uma eficiente solução para outros riscos existenciais, e, erroneamente, estaremos ignorando nossa melhor oportunidade de sobrevivência. Subestimar o potencial impacto da Inteligência Artificial significa subestimar tantos os impactos positivos quanto negatives. Por isso o título desse capítulo é “A Inteligenbcia Artificial como fator positive e negative no risco global”, e não “Riscos Globais da Inteligência Artificial”. O prospecto de AI interage com o risco global de maneiras mais complexas que essa, fosse a AI só problemas, a questão seria trivial.

4: Capacidade e Motivação

Na discussão sobre AI há uma falácia normalmente cometida, principalmente quando se trata de capacidade sobrehumana. Alguém diz:

”Quando a tecnologia avançar suficientemente, seremos capazes de criar mentes muito mais inteligentes que as humanas. Mas, é óbvio que quanto maior nossa inteligência maior será a pizza que conseguiremos fazer. Uma superinteligência poderia fazer uma pizza super grande, do tamanho de uma cidade. Santo Deus! O futuro será infestado de pizzas gigantes!” A questão aqui é se uma superinteligência vai querer fazer pizzas gigantes. A visão salta diretamente de capacidade para factualidade sem considerer a necessidade de motivação.

As seguintes linhas de raciocínio, consideradas isoladamente sem argumentos sólidos, exibem, sem exceção a Falácia da Pizza Gigante:

Uma Inteligência Artificial suficientemente poderosa iria sobrepujar qualquer resistência humana e dizimar toda a humanidade. (e a AI iria querer fazer isso). Portanto não devemos construir AIs.

Uma Inteligência Artificial suficientemente poderosa seria capaz de desenvolver novas tecnologias médicas e salvarmilhões de vidas humanas. (e a AI iria querer fazer isso). Portanto deveríamos construir AIs.

Uma vez que os computadores tornam-se baratos o suficiente, a vasta maioria dos trabalhos seriam executáveis por Inteligência Artificial mais facilmente que por humanos.

Uma Inteligência Artificial suficientemente poderosa seria melhor que nós em matemática, engenharia, música, arte e em todas outras tarefas que considerarmos relevantes. e a AI iria querer fazer isso). Portanto, após a criação da AI os humanos terão nada para fazer, e morreremos de fome de tanto assistir televisão.

4.1 Processos de Otimização

A deconstrução acima da Falácia da Pizza Gigante invoca um antropomorfismo intrínseco, a idéia que as motivações são separáveis: a suposição implícita que, por falar de capacidade e motivação como entidades separadas, nós estamos cortando a realidade exatamente nas suas juntas. Esta é uma fatia útil, mas ainda uma fatia antropomórfica.

Para visualizar a questão em termos gerais, vou introduzir o conceito de um processo de otimização: um sistema que acerta alvos pequenos em grandes espaços de busca a fim de produzir efeitos coerentes no mundo real.

Um processo de otimização dirige o futuro a uma região particular dentro do que é possivel. Estou visitando uma cidade distante, um amigo local se oferece para me levar até o aeroporto. Eu não conheço o caminho. Quando meu amigo chega a um cruzamento, eu não sei dizer se ele vai virar à

direita ou à esquerda, nem individual nem sequencialmente. Ainda assim posso prever o resultado da ação de meu amigo: nós vamos chegar ao aeroporto. Mesmo se a casa dele fosse localizada em outro ponto da cidade, fazendo com que ele fizesse uma sequência completamente diferente, I poderia prever com segurança nosso destino. Não é estranho se encontrar nessa situação, cientificamente falando? Eu posso prever o resultado de um processo sem ser capaz de prever nenhum de seus passos intermediários. Falarei agora da region para a qual o processo de otimização guia o future como o alvo do otimizador.

Considere um automável, digamos um Toyota Corolla. De todas as combinações possíveis para os átomos que constituem um Corolla, somente uma fração infinitesimal se qualifica como um automóvel functional. Se você junta as moléculas ao acaso, muitas e muitas eras passarão antes que você obtenha um carro. Uma pequena fração do espaço “design” decreve veículos que poderíamos reconhecer comomais rápido, eficiente e seguro que o Corolla. Portanto o Corolla não é ótimo dentro dos objetivos do designer. O Corolla é, contudo, otimizado, pois o designer teve que atingir um alvo comparativamente infinitesimal no espaço “design” a fim de criar um carro functional, sem falar de um carro com as qualidades de um Corolla. Não se consegue construir uma carruagem functional cortando tábuas ao acaso e aparafusando na base do cara ou coroa. Para atingir alvos tão pequenos no espaço “configuração” requer um poderoso processo de otimização.

A noção de um “processo de otimização” é previsivemente útil, pois pode ser mais fácil compreender o alvo de um processo de otimização que sua dinâmica passo a passo. A discussão acima sobre o Corolla assume implicitamente que o designer do Corolla estava tentando produzir um veículo, um meio de transporte. Esse pressuposto deve ficar evidente, mas não é errado e ajuda muito na compreensão do Corolla.

4.2 Apontando para o Alvo

A tentação é de se perguntar: “O que as AIs vão querer.” Esquecendo-se que o espaço das mentes em geral é muito mais vasto que o pequeno ponto humano. Deveríamos resistir à tentação de espalhar quantificadores sobre todos os tipos de mentes possíveis. O contador de estórias, ditando contos da distante e exótica terra chamada Futuro, dirá como será esse futuro. Eles fazem previsões. Eles dizem: “AIs vão atacar os humanos com um excército de robôs” ou “AIs descobrirão a cura do câncer”. Eles não propõe relações complexas entre as condições iniciais e os resultados, isso iria afujentar a audiência. Mas nós precisamos de compreensão relacional a fim de manipular o futuro, guiá-lo à uma região favorável à humanidade. Se não guiarmos, corremos acabar onde estamos indo.

O desafio não é prever o que as AIs farão, a tarefa não é nem fazer previsões para um design particular de AI. Na verdade, a missão é escolher um processo de otimização particularmente poderoso, cujo benefícios possam ser legitimamente avaliados.

Eu peço encarecidamente aos meus leitores para não começar a buscar razões pelas quais um processo de otimização totalmente genérico seria amigável. A seleção natural não é amigável, nem lhe odeia e nem lhe deixará em paz. A evolução não pode ser antopromorfizada, ela não funciona da mesma forma que você. Muito biologistas prédecada de 1960 esperavam que a seleção natural faria todo o tipo de coisas belas, e racionalizaram todo tipo de razões elaboradas pelas quais ela faria dessa forma. Eles ficaram desapontados, pois a própria seleção natural não começou sabendo que ela queria um resultado humanamente belo, e então racionalizaram caminhos elaborados a fim de produzir belos resultados usando a pressão da seleção. Dessa forma os eventos na natureza foram produtos de diferentes processo causais do que ia na mente dos biologistas pré 60', de forma que a previsão e a realidade se divergiam. Wishful thinking adiciona detalhes, constringe previsões, e cria, portanto, um fardo de improbabilidade. O que dizer do engenheiro civil que tem a esperança de a ponte não cair? Deveria ele argumentar que pontes, em geral, não possuem propensão a cair? Mas a natureza não racionaliza as razões pelas quais as pontes não deveriam cair. Ao contrário, o engenheiro se livra do fardo da improbabilidade atraves escolhas específicas guiadas por compreensões específicas. O engenheiro civil começar por desejar uma ponte, usa então uma teoria rigorosa escolher o projeto de uma ponte que suporta carros; Constroi enfim uma ponte no mundo, cuja estrutura reflete o design calculado, e assim a estrutura no mundo real suporta carros, obtendo portanto harmonia entre a previsão dos resultados positivos e e a realização dos mesmos.

5: A AI Amigável

Seria uma boa coisa se a humanidade soubesse como escolher e criar um processo de otimização poderoso com um alvo em particular. Ou em termos mais colloquial, seria bom se nós soubéssemos como construir um bom AI.

Para descrever o campo de conhecimento necessário para trabalhar nesse desafio, eu propus o termo "AI Amigável". Além de me referir a um conjunto de técnicas, "AI Amigável" também pode se referir ao produto da técnica, um AI criado com motivações específicas. Quando uso o termo Amigável em qualquer sentido, eu o capitalizo para evitar confusão com o intuitivo senso de "amigável".

Uma reação comum que encontro é as pessoas imediatamente declarar que AI amigável é um sonho impossível, pois uma AI suficientemente poderosa será capaz de alterar seu próprio código fonte a fim de quebrar qualquer limite colocado nela.

O primeiro defeito que você deveria notar é a Falácia da Pizza Gigante. Qualquer AI com livre acesso ao seu próprio código fonte iria, em princípio, modificá-lo de forma a alterar seu alvo de otimização. Isso não implica que a AI tem a motivação para alterar suas próprias motivações. Eu não iria conscientemente engolir uma pílula que me faria ter prazer em cometer assasinatos, pois atualmente eu prefiro que meus companheiros humanos não morram.

Mas e se eu tentar me alterar e cometer um erro? Quando os engenherios computacionais testam um chip, uma boa idéia se o chip tem 155 milhões de transistores e você não pode remendá-los depois, os engenheiros usam prova formal verificado por máquina, mas guiado por humanos. Um ponto glorioso sobre a prova matemática é que uma prova de 10 milhões de passos é tão confiável quanto outra de 10 passos. Mas o ser humano não é tão confiável para se debruçar sobre uma prova de 10 milhões de passos; teremos uma chance muito alta de não perceber um erro. E as autais técnicas para provar teoremas não são eficientes a ponto de desenhar e provar, por si só e completamente, um chip de computer, os algorítimos atuais sofrem uma explosão exponencial no espaço de busca. Matemáticos humanos podem provar teoremas muito mais complexos que os modernos algorítimos padem manejam sem sofre tal explosão. Mas a matemática humana é informal e infiel; ocasionalmente alguém encontra uma falha numa prova de teorma aceita previamente. A saída está em os engenheiros humanos guiarem o provador de teoremas através dos passos intermediaries de uma prova. O humano escolhe o próximo lema, e um provador de teoremas complexo gera a proval formal e um simples verificador checa os passos. É dessa forma que os engenheiros modernos constroem mecanismos seguros com 155 milhões de partes independentes.

Para testar um chip de computador de forma correta é necessário uma sinergia entre a inteligência humana e algotítimos de computador, pois por enquanto nenhum deles é autosuficiente. Talvez uma AI de verdade poderia usar uma combinação similar de habilidades quando for modificar seu próprio código fonte, ou seja, teria ambas capacidades de inventar grandes design sem sofrer uma explosãoexponencial e de verificar seus passos com extrema confiabilidade. Esse é certamente um caminho para que uma verdadeira AI se mantivesse estável em seus objetivos, mesmo depois de levar a cabo um grande número de automodificações.

Esse *paper* não ira explorer a idéia acima em detalhes. (Mas vjea Schmidhuber 2003 para uma noção relacionada). But devemos pensar sobre o desafio e estudá-lo nos melhores detalhes técnicos disponíveis antes de declará-lo impossível, especialmente se o algo muito relevante depende dessa resposta. É falta de respeito à engenhosidade humana declarar um desafio como impossível sem estudá-lo em detalhes exercendo a criatividade. Dizer que você não pode fazer algo é uma declaração muito

pesada, que você não pode construer uma máquina voadora mais pesada que o ar, que você não pode obter energia útil a parir de reações nucleares, que você não pode voar até a lua. Tais declarações são generalizações universais, quantificadas sobre cada tentativa que alguém fez ou fará em prol da resolução do problema. Precisamos apenas de um simple contraexemplo para desqualificar um quantificador universal. A afirmação que uma AI amigável é teoricamente impossível, ousa quantificar sobre cada design de mente possível e cada processo de otimização possível, incluindo seres humanos, que também são mentes, alguns dos quais são amáveis e desejam que eles sejam ainda mais amáveis. Nesse ponto há uma série de razões plausíveis pelas quais uma AI Amigável poderia ser humanamente impossível, e é ainda mais provável que o problema é solucionável, mas ninguém chegará à solução em tempo. De qualquer forma não devemos descartar essa possibilidade tão rapidamente, especialmente se considerarmos o que está em jogo.

6: Falhas Técnicas e Falhas Filosóficas

Bostrom (2001) define como catastrophe existencial aquela na qual a vida inteligente originada na Terra é destruida em caráter permanente ou a que destroi parte de seu potencial. Podemos divider as falhas potenciais das tentativas de se criar uma AI Amigável em duas categorias informais de fuzzy, falha técnica e falha filosófica. Falha técnica é quando você tenta construir uma AI, mas a coisa não funciona do modo esperado, você falhou em compreender o funcionamento de seu próprio código fonte. Falha filosófica é tentar construer a coisa errada, de forma que mesmo que você obtenha êxito, ainda assim falhará em ajudar qualquer pessoa ou beneficiar a humanidade. Não carece dizer, mas as duas falha não são mutuamente exclusivas.

A linha que divide esses dois casos é tênue, pois a maioria das falhas filosóficas são muito mais fáceis de explicar quando na presença de conhecimento técnico. Na teoria, você deve primeiro dizer o que você deseja, então pensar numa maneira de obtê-lo. Na prática é necessário um profundo conhecimento técnico para pensar e decider o que você de fato quer.

6: Falhas Técnicas e Falhas Filosóficas

No final do sec. XIX, muitas pessoas honestas e inteligentes defendiam o comunismo, na melhor das boas intenções. As pessoas que incialmente inventaram, espalharam e engoliram o meme comunista eram, historica e de fato , idealistas. Os primeiros comunistas não tinham o exemplo da Rússia Soviética para alertá-los. Na época, sem o benefício de hindsight, parecia uma idéia muito boa, Após a revolução, quando os comunistas ganharam o poder e foram corrompidos por ele, outras motivações podem

ter dominado a cena; mas isso não era o que os primeiros idealistas previram, previsível ou não. É importante compreender que os autores de grandes catastrophes não são necessariamente maus nem inalbitualmente estúpidos. Se atribuirmos todas as tragédoas ao mal ou à estúpidez, nós olharemos a nós mesmos, corretamente perceber que não somos maus ou estúpidos e dizer: “Mas isso nuca aconteceria conosco”.

O que os primeiros revolucionários comunistas pensavam que aconteceria, como conseqüência empírica de sua revolução, era que as vidas das pessoa melhorariam: que os operários teriam uma jornada mais curta em trabalhos menos cansativos e seriam ainda mais bem remunerados. Colocando de forma branda, essa expectativa, no entanto, não se confirmou. Mas, o que o pensamento desses primeiros comunistas previa, não era assim muito diferente do previsto pelo pensamento dos defensores de outros sistemas politicos, como conseqüência empírica de seus sistemas politicos favoritos. Pensaram que as pessoas ficariam felizes. Eles estavam enganados.

Imagine agora que alguém tente programar uma AI “amigável” para impplementar o comunismo, ou o libertarianismo, ou o anarco-feudalismo, ou o seu sistam politico favorite qualquer, acreditando que isto fará da utopia uma realidade. Os sistemas políticos favoritos das pessoas inspiram expectativas altamente positivas, de forma que a proposta soará como uma idéia realmente boa àqueles que a propõem.

Nós poderíamos ver a falha do programador em um nível moral ou ético - diga que é o resultado de alguém que confía demasiadamente em si mesmo, deixando de considerar sua própria falibilidade, recusando-se a considerar a possibilidade que o comunismo, depois de tudo, poderia estar equivocado. Mas na línguagem da teoria Bayesian da decisão, há uma visão técnica complementar do problema. Da perspectiva da teoria da decisão, a escolha do comunismo resulta da combinação de uma opinião empírica com um julgamento do valor. A opinião empírica é que, o comunismo, quando implementado, resulta em uma realidade específica ou em tipos de realidades específicas, ou seja, as pessoas ficarão mais satisfeitas, trabalharão poucas horas, e possuirão uma riqueza material maior. Esta é definitivamente uma previsão empírica; mesmo a parte sobre a felicidade é , de fato, uma propriedade dos estados do cérebro, contudo difícil de medir. Se você implantar o comunismo, esse resultado pode ou não se concretizar. O julgamento do valor é que este resultado é satisfatório ou é preferível às circunstâncias atuais. Dado uma opinião empírica diferente sobre as conseqüências reais de um sistema comunista, a decisão pode submeter-se a uma mudança correspondente.

Nós esperaríamos que um AI verdadeiro, uma inteligência artificial geral, fosse capazes de mudar sua opinião empírica. (Ou seus modelos

probabilísticos do mundo real, etc.). Se de algum modo Charles Babbage tivesse vivido antes de Nicolaus Copernicus, e os computadores fossem inventados antes dos telescópios, e os programadores desses dias criassem com sucesso uma inteligência artificial geral, não aconteceria que a AI acreditaria para sempre que o sol orbitou a terra. O AI poderia transcender o erro factual de seus programadores, contanto que os programadores entendessem mais de inferência que de a astronomia. Para construir um AI que descubra as órbitas dos planetas, os programadores não necessitam saber amatemática da mecânica Newtoniana, somente a matemática da teoria de probabilidade Bayesiana.

A falha em programar uma AI para implantar o comunismo, ou qualquer outro sistema político, é que você está programando meios em vez de fins. Você está programando em uma decisão fixa, sem que essa decisão seja reajustável após ter adquirido o conhecimento empírico melhorado sobre os resultados do comunismo. Você está dando à AI uma decisão fixa sem dizer a ela para reavaliar, em um nível mais elevado da inteligência, o processo falível que produziu essa decisão.

Se eu jogar xadrez contra um jogador mais forte, eu não posso predizer exatamente qual seu próximo lance - se eu podesse faze-lo, eu seria necessariamente no mínimo tão forte quanto ele. Mas eu posso prever o resultado do jogo, que é uma vitória do meu oponente. Eu vejo os futuros possíveis para a qual ele está mirando, fato que me permite prever o destino final, mesmo se eu não puder ver o trajeto. Quando eu estou no meu momento mais creativo, é mais difícil predizer minhas ações, e mais fácil de predizer as conseqüências de minhas ações. (desde que eu conheça e compreenda meus objetivos!) Se eu quiser um jogador de xadrez melhor que um humana, eu tenho que programar uma busca para movimentos vencedores. Eu não posso programar movimentos específicos porque então o jogador não sera melhor do que eu. Quando eu lanço uma busca, eu sacrifico necessariamente minha abilidade de predizer a resposta exata. Para obter uma boa resposta, deve-se sacrificar a habilidade de predizê-la, não apesar da habilidade de se fazer a pergunta.

Tal confusão como programar o comunismo diretamente, provavelmente não atraia um programador de AI geral que falasse a línguagem da teoria da decisão. Eu chamaria isso de falha filosófica, mas, podemos culpar a falta do conhecimento técnico.

6.2: Um exemplo de falha técnica

“No lugar das leis que limitem o comportamento de máquinas inteligentes, nós necessitamos dar-lhes emoções que possam guiar sua aprendizagem de comportamentos. Devem nos querer felizes e prósperos, que é a emoção que nós chamamos o amor. Nós podemos projetar máquinas inteligentes de modo que sua emoção preliminar, nativa é o amor incondicional para todos os seres humanos. Primeiramente nós podemos construir máquinas relativamente simples que aprendem a reconhecer a

felicidade e a tristeza em expressões facial humanas, em vozes humanas e na linguagem do corpo humano. Então nós podemos colocar o resultado deste aprendizado como valores emocionais nativos em máquinas inteligentes mais complexas, reforçados positivamente quando nós estamos felizes e reforçados negativamente quando nós estamos infelizes. As máquinas podem aprender algoritmos para predizer o futuro de forma aproximada, como por exemplo os investores usam atualmente máquinas que aprendem a predizer os preços de títulos futuros. Assim nós podemos programar máquinas inteligentes para aprender algoritmos para prever a felicidade humana, e usar essas previsões como valores emocionais. "[2]

O texto abaixo, embora famoso e apesar de ser normalmente considerado fato verdadeiro, podem ser apócrifo. Eu não encontrei um relato de primeira mão. Para relatos não referenciados vide, por exemplo, Crochat e Franklin (2000) ou http://neil.fraser.name/writing/tank/. Entretanto, as falhas do tipo são uma preocupação principal do mundo real ao construir e ao testar redes neurais.[3]

Bill Hibbard, após ter visto um esboço deste paper, escreveu uma resposta que discute que a analogia ao problema do "classificador de tamques" não se aplica ao reforço do aprendizado em geral. Sua crítica pode ser encontrada no link http://www.ssec.wisc.edu/~billh/g/AIRisk_Reply.html . Minha resposta pode está no http://yudkowsky.net/AIRisk_Hibbard.html . Hibbard anota também que a proposta de Hibbard (2001) foi substituída por Hibbard (2004). O último recomenda um sistema de duas camadas em que as expressões de aceitação dos seres humanos reforcem o reconhecimento de felicidade, que uma vez reconhecida, reforce as estratégias da ação. -- Bill Hibbard (2001), Máquinas Super-inteligentes.

Uma vez, o exército dos E.U. temtou usar redes neurais para detectar automaticamente os tanques inimigos sob camuflagem. Os investigadores treinaram uma rede neural com 50 fotos dos tanques camuflados com folhagem, e 50 fotos de folhagem sem tanques. Usando técnicas padrão para a aprendizagem supervisionada, os investigadores treinaram a rede neural de forma que o resultado foi 100% correto. Isto, contudo não assegurou, nem mesmo implicou, que novos exemplos seriam classificados corretamente. A rede neural pode "ter aprendido" 100 casos especiais que não generalizariam a nenhum problema novo. Sabiamente, os investigadores tinham feito exame prévio de 200 fotos, 100 de tanques e 100 de árvores. Tinham usado somente 50 de cada um para o jogo do treinamento. Os pesquisadores rodaram a rede neural nas 100 fotos restantes, e sem treinamento adicional a rede neural classificou todas as fotos restantes corretamente. Sucesso confirmado! Os investigadores entregaram o trabalho terminado ao Pentágono, que o devolveu em seguida queixando-se que em seus próprios testes a rede neural fêz não melhor do que umaescolha ao acaso.

Após uma revisão descobriu-se que, na série de dados dos pesquisadores, as fotos dos tanques camuflados tinham sido feitas em dias nublado, quando as fotos da floresta tinham sido feitas em dias ensolarados. A rede

neural tinha aprendido distinguir dias nublados dos ensolarados ensolarados, em vez de distinguir os tanques camuflados de floresta sem tanques[2].

Uma falha técnica ocorre quando o código não faz o que você pensa que ele faz, embora execute fielmente como você o programou. Mais de um modelo podem carregar os mesmos dados. Suponha que nós treinássemos uma rede neural para reconhecer as caras humanas sorrindo para as distinguir das caras humanas contrariadas. A rede classificaria um retrato minúsculo de um *smiley* como uma cara humana sorrindo? Se um AI ligado a tal código possuísse o poder, e Hibbard (2001) falou de superinteligência, então nossa galáxia terminaria entupida de fotos moleculares de *smileys*?[3]

Este tipo de falha é especialmente perigoso porque parecerá funcionar dentro de um contexto fixo, falha então quando o contexto muda. Os pesquisadores da história do “classificador de tanques” ajustaram a rede neural até que carregou corretamente os dados do treinamento, a seguir verificaram a rede em dados adicionais (sem mais mais ajustes). Infelizmente, os dados do treinamento e os dados da verificação pareceram compartilhar uma suposição sobre todos os dados usado no desenvolvimento, mas não sobre todos os dados disponíveis no mundo rea. Na história do classificador do tanque, a suposição é que os tanques foram fotografados em dias nublados.

Suponha que nós desejássemos desenvolver uma AI com poder crescente. O AI estánum estágio de desenvolvimento onde os programadores humanos sejam mais poderosos do que ela - não no sentido do mero controle físico sobre a fonte elétrica do AI, mas no sentido que os programadores humanos são mais espertos, mais criativo, mais perspicaz que a AI. Durante o período de desenvolvimento nós supomos que os programadores possuam a habilidade de fazer mudanças ao código de fonte da AI sem necessitar o consentimento da AI. Entretanto, pretende-se que a AI tenha os estágios posdesenvolvimento, incluindo, no exemplo do cenário de Hibbard, inteligência subrehumana. Um AI com essa inteligência certamente não poderia ser modificado sem seu consentimento. Neste momento nós devemos confiar que o sistema de objetivo previamente colocado irá funcionar corretamente, porque se ela operar de uma forma suficientemente imprevisível, a AI pode ativamente resistir a nossas tentativas de a corrigir - e, se elaI for mais esperta do que um ser humano, provavelmente nos vencer.

Tentar controlar uma AI de poder crescente treinando uma rede neural para fornecer seu próprio sistema de objetivos enfrenta o problema de uma mudança enorme de contexto entre o estágio de desenvolvimento da AI e o de posdesenvolvimento. Durante o estágio de desenvolvimento, o AI pode somente produzir os estimulos que caem na categoria das “caras humanas

sorrindo", resolvendo tarefas humanas fornecidas, como seus fabricantes pretenderam. Vejamos o futuro numa época quando a AI é inteligente de maneira sobre humana e construiu sua própria infraestrutura nanotecnoçógica, e a AI pode produzir os estimulos classificados no mesmo rotulador enchendo a galáxia de minúsculos *smilyes.*

Dessa forma a AI funciona perfeitamente no estágio de desenvolvimento, mas produz resultados catastróficos depois de se tornar mais inteligente que seus programadores.

Há uma tentação de pensar, "mas certamente a AI saberá que não é o que nós pretendíamos?" Mas o código não está dado à AI, para que ela perceba o erro e corrija. O código é a AI. Talvez com bastante esforço e compreensão nós podemos escrever o código que se importa se nós escrevermos o código errado - a instrução legendária de FOQUEP, que entre programadores está para "faça o que eu pretendo". (Raymond 2003.) Mas o esforço é requerido para escrever um FOQUEP dinâmico, e em nenhuma parte de Hibbard está a menção de projetar uma AI que faça o que nós pretendemos, não o que nós dizemos. Os chips modernos não tem FOQUEP em seu código; não é uma propriedade automática. E se você alterar seu FOQUEP próprio, você sofreria as conseqüências. Por exemplo, suponha que o FOQUEP foi definido como maximar a satisfação do programador com o código; quando o código executado como uma superintelligence, ele poderia reescrever o cérebro do programador para ficar extremamente satisfeitos. Eu não digo que isso é inevitável; Eu indico somente que o FOQUEP é o desafio técnico principal e complexo do "AI Amigável".

7: Taxas de crescimento da inteligência

Do ponto de vista do risco existential, um dos pontos mais críticos sobre a inteligência artificial é que ela pode aumentar sua própria inteligência de forma extremamente rápida. A razão óbvia para suspeitar desta possibilidade é automelhoria recursiva. (Good 1965) A AI torna-se mais esperta, includindo tornar-se mais esperta na tarefa de escrever as funções cognitivas internas de uma AI, assim que ela pode reescrever suas funções cognitivas existentes para trabalhar realmente melhor, que faz de si ainda mais esperta, includindo mais esperta na tarefa de se reescrever, e assim por diante de modo exponencial.

Os seres humanos não se automelhoram recursivamente de maneira significativa. A uma extensão limitada, nós nos melhoramos: nós aprendemos, nós praticamos, nós afiamos nossos habilidades e conhecimento. A uma extensão limitada, estas automelhorias melhoram nossa abilidade de melhorar. As descobertas novas podem aumentar nossa abilidade de fazer umas descobertas adicionais - nesse sentido, o conhecimento se alimenta em si mesmo. Mas há ainda um níve subjacente nós não tocamos ainda. Nós não reescrevemos o cérebro humano. O cérebro é, definitivamente, a fonte da descoberta, e nossos cérebros são hoje exatamente como eram há dez mil anos atrás.

Em um sentido similar, a seleção natural melhora organismos, mas o processo da seleção natural propriamente dito não melhora - não em um sentido forte. A adaptação pode abrir o caminho para adaptações adicionais. Neste sentido, a adaptação alimenta- em si mesma. Mas se a panela de genes ferve, é porque ainda há um aquecedor por perto, que é o processo de mutação, recombination e seleção, que não é re-arquitetado autonomamente. Algumas inovações raras aumentaram a taxa da evolução, tal como a invenção da recombinação sexual. Mas mesmo o sexo não mudou a natureza essencial da evolução: sua falta da inteligência abstrata, sua dependência em mutatções aleatórias, sua cegueira e incrementalismo, seu foco em freqüências dos alelos. Similarmente, nem mesmo a invenção da ciência mudou o caráter essencial do cérebro humano: seu núcleo límbico, seu cortex cerebral, seus automodelos prefrontais nem sua velocidade característica de 200Hz.

Uma inteligência artificial poderia reescrever seu próprio código, poderia mudar a dinâmica subjacente de otimização. Tal processo de otimização seria muito mais eficiente do que a evolução que acumula adaptações, ou dos seres humanos que acumulam o conhecimento. A implicação chave para nossas finalidades é que um AI poderia fazer um salto enorme na inteligência após ter alcançado algum ponto inicial crítico.

Se encontra frequentemente o ceticismo sobre este cenario, que Good (1965) chamou da “uma explosão de inteligência”, pois o progresso na inteligência artificial tem a reputação de ser muito lento. Neste momento pode provar útil rever uma surpresa histórica de certa forma análoga. (O que segue é tomado principalmente de Rhodes 1986.)

In 1933, o Lorde Ernest Rutherford disse nunca seria possível extrair energia a partir de fissão nuclear: “ Qualquer um que busque uma fonte de energia na transfomação do átomo pode ser chamado de lunatic”. Naquela época semanas e horas de trabalho duro eram necessaries para dividir um átomo.

Alguns anos depois, em 1942, em uma quadra de squash de Stagg Field na universidade de Chicago. Os físicos estavam construindo algo em forma maçaneta gigante a partir de camadas alternadas de grafite e de urânio, e pretendia começar a primeira reação nuclear autosustentável. Encarregado do projeto estava Enrico Fermi. O número chave para a pilha é k, o fator eficaz da multiplicação do nêutron: o número médio dos nêutrons de uma reação de fissão que causa uma outra reação de fissão. Em k menos um, a pilha é subcritical. No k maior ou igual a 1, a pilha deve sustentar uma reação crítica. Fermi calcula que a pilha alcançará k = 1 entre as camadas 56 e 57.

Um grupo do trabalho conduzido por Herbert Anderson termina a camada 57 na noite de 1º de dezembro de 1942. As hastes de controle, tiras de madeira cobertas com folha de cádmio absorvente de neutrons, impedem a pilha de alcançar o ponto crítico. Anderson remove tudo com exceção de uma haste de controle e mede a radiação da pilha, confirmando que a pilha está pronta para uma reação em cadeia no dia seguinte. Anderson introduz

todas as hastes do cádmio e trava-as no lugar com cadeados, a seguir fecha a quadra de squash e vai para casa.

No dia seguinte, 2 de dezembro de 1942, em uma manhã ventosa de temperaturas abaixo de zero, Fermi começa a experiência final. Todas com exceção de uma das hastes de controle são retiradas. As 10:37 am, Fermi requisita a retirada da haste de controle até a metade. Os contadores de Geiger estalam mais rapidamente, e a caneta do gráfico move-se para cima. “Não está correto,” diz Fermi, “o traço vai atingir esse ponto e parar de subir,” indicando um ponto no gráfico. Em alguns minutos o gráfico vem ao ponto indicado e para. Sete minutos mais tarde, Fermi pede para retirar a haste um pouco mais. Outra vez a radiação aumenta e para. A haste é puxada para fora outras seis polegadas, então outras e outras. Às 11:30, a ascensão lenta do gráfico é parade por um enorme CRASH, uma haste de controle da emergência, acionada por uma câmara de ionização, é ativada e fecha a pilha, que está ainda aquém do ponto crítico. Fermi calmamente pede a equipe para fazer um break para o almoço.

Às 2pm a equipe recomeça, retira e trava a haste de controle de emergência, e move a haste de controle para seu último ajuste. Fermi faz algumas medidas e cálculos, a seguir começa outra vez o processo de retirar a haste em incrementos lentos. Em 3:25 pm, Fermi pede para retirar outras doze polegadas. “Acho que vai funcionar”, Fermi diz. “Agora vai ficar autpsustentável. O traço não vai parar de subir. “

Herbert Anderson reconta (de Rhodes 1986):

“Inicialmente você podia ouvir o som do contador do nêutron, clickety-clack, clickety-clack. Então os cliques vieram mais e mais rapidamente, e após um tempo começaram a se fundir em um rugido; o contador não podia seguir mais. Aquele era o momento para mudar para o registrador de gráfico. Mas, quando a troca foi feita, fez-se silêncio repentino na sala ao notarem que a caneta de reistro havia-se entortado. Era um silêncio estranho. Todos sabiam da importância dessa troca; nós estávamos no regime de intensidade elevada e os contadores eram incapazes de continuar lidando com a situação. Repetidas vezes, a escala do registrador teve que ser mudada para acomodar a intensidade do nêutron que estava aumentando mais e mais rapidamente. De repente Fermi levantou sua mão. “A pilha ultrapassou o ponto crítico” anunciou. Ninguém presente teve a dúvida sobre isso.”

Fermi manteve a pilha funcionando por 28 minutos, com a intensidade de nêutron dobrando a cada cada dois minutos. A primeira reação crítica teve k de 1.0006. Mesmo em k=1.0006, a pilha era somente controlável porque alguns dos nêutrons de uma reação de fissão do urânio são atrasados - vêm da deterioração de subprodutos de vida curta da fissão. Para cada 100 fissões em U235, 242 nêutrons são emitidos quase imediatamente

(0.0001s), e 1.58 nêutrons são emitidos numa média dez segundos de mais tarde. Assim a vida média de um nêutron é ~0.1 segundos, implicando 1.200 gerações em dois minutos, e dobrando a cada dois minutos porque 1.0006 elevado à potencial de 1.200 é ~2. Uma reação nuclear é crítica sem a contribuição de nêutrons atrasados. Se a pilha de Fermi fosse dada como crítica com k=1.0006, a intensidade do nêutron dobraria a cada décimo de um segundo.

A primeira moral da história é que confundir a velocidade da pesquisa da AI com a velocidade de uma AI real depois de construída é como confundir a velocidade da pesquisa da física com a velocidade de reações nucleares. Mistura o mapa com o território. Levou-se anos para construir a primeira pilha, trabalho de um grupo pequeno de físicos que não gerou muita mídia. Mas, uma vez que a pilha foi construída, as coisas interessantes aconteceram no na escala de tempo das interações nucleares, não na escala do discurso humano. No domínio nuclear, as interações elementares acontecem muito mais rapidamente do que nos neurônios humanos. Muito do mesmo pode ser dito dos transistores.

Uma outra moral é que há uma diferença enorme entre uma automelhoria com fator de 0,9994 e outra de 1.0006. A pilha nuclear não cruzou o ponto crítico porque a equipe empilhou mais material. Os físicos empilharam de maneira lenta e regular. Mesmo se houver uma curva subjacente suave da inteligência do cérebro em função da pressão da otimização exercida previamente nesse cérebro, a curva da automelhoria recursiva pode mostrar um salto enorme.

Há também outras razões porque uma AI poderia mostrar um repentino e grande salto na inteligência. Os homo sapiens mostraram um salto relvante na eficácia da inteligência, como o resultado da seleção natural exercendo mais-ou-menos a pressão constante da otimização em hominídeos por milhões de anos, gradualmente expandindo o cérebro e o cortex prefrontal, mexando na a arquitetura do software. Há algumas dezenas de milhares de anos atrás, a inteligência hominídea cruzou algum ponto chave e fez-se um salto enorme na eficácia no mundo real; nós saímos das cavernas aos arranhacéus num piscar de um olho em termos evolucionários. Isto aconteceu com uma pressão subjacente contínua da seleção, não havia um salto enorme no poder da otimização da evolução quando os seres humanos vieram. A arquitetura subjacente do cérebro era também contínua - nossa capacidade cranial não dobrou aumentou repentinamente. Assim pode ser que, mesmo se a AI fosse elaborada programadores humanos, a curva para a inteligência eficaz saltará de agudamente.

Ou talvez alguém construa um protótipo do AI que mostre alguns resultados prometedores, e o programa demostrativo atraia outros USD 100 milhões como capital de risco, e compra-se com este dinheiro mil vezes mais poder de supercomputação. Eu duvido que um aumento na ordem de milhares no hardware compraria qualquer coisa com um aumento da ordem de milhares em inteligência eficaz - mas a mera dúvida não é de confiança na ausência de habilidade para executar um cálculo analítico. Comparado aos chimpanzés, os seres humanos têm uma vantagem três vezes maior no

cérebro e uma vantagem de seis vezes no cortex prefrontal, que sugere (a) que o software é mais importante do que o hardware e (b) que aumentos pequenos no hardware pode suportar melhorias grandes no software. É um ponto a mais a considerar.

Finalmente, a AI pode fazer um salto aparentemente relevante na inteligência puramente como o resultado do anthropomorfismo, a tendência humana no “do idiota vila” e “Einstein” como as extremidades extremas da escala da inteligência, em vez dos pontos quase indistinguiveis na escala da mente-em-geral. Qualquer coisa mais idota do que um ser humano idiota pode aparecer-nos como simplesmente “idiota”. Pode-se imaginar a inteligência das AIs movendo-se firmemente para cima na escala da inteligência, movendo-se além da dos ratos e chimpanzés, mas ainda como uma AI idiota, pois AIs não podem falar a língua fluente ou escrever papers de ciências, e então inteligência da AI cruza a abertura minúscula do infra-idiot ao ultra-Einstein no curso de um mês ou de algum período similarmente curto. Eu não penso que este cenario exato é plausivel, na maior parte porque eu não espero a curva da automelhoria recursiva se mover em num ritmo rastejamento linear. Mas eu não sou o primeiro para indicar que a “AI” é um alvo movel. Assim que um marco for conseguido realmente, cessa de ser “AI”. Isto pode somente incentivar o adiamento.

Para fins de argumentação, vamos conceber que, por tudo que indica (e me parece também probavel no mundo real) que um AI tem a potencialidade para fazer um pulo repentino, nítido e grande na inteligência. O que acontece depois?

Primeiramente: segue que, uma reação que eu me ouço frequentemente, “não necessitamos nos preocupar sobre a AI amigável porque nós não temos ainda a AI”, é um pensamento equivocado ou simplesmente suicida. Nós não podemos confiar em ter o aviso avançado distante antes que a AI esteja criada; os avanços tecnológicos do passado não se telegrafaram às pessoas vivas da época, somos profetas do passado. A matemática e as técnicas da AI amigável não irão se materializar do nada. São necessários anos para fazer fundações firmes. E nós necessitamos resolver o desafio da AI Amigável antes que a inteligência artificial geral esteja criada, não depois; Eu não deveria

> nem mesmo ter que apontar esse detalhe. Haverá umas dificuldades para desenvolver a AI amigável porque o próprio campo da AI está em um estado de consenso baixo e entropia elevada. Mas isso não significa que nós não necessitamos se preocupar sobre o AI amigável. Significa que haverá umas dificuldades. As duas indicações, infelizmente, não são nem remotamente equivalentes.

A possibilidade de grandes afiados na inteligência implica também num padrão mais elevado de técnicas para criar AI Amigáveis. A técnica não pode supor a habilidade dos programadores de monitorar a AI de contra a sua vontade, e de reescrever a AI contra sua vontade, chama à possibilidade de uma ameaça military superior; nem pode o algoritmo supor que os programadores controlam da uma tecla “recompensa” que uma AI mais esperta poderia arrancar dos programadores; et cetera. Certamente

ninguém devem, para começar, a fazer estas suposições. A proteção indispensavel é um AI que não queira lhe ferir. Sem o indispensavel, nenhuma defesa auxiliar pode ser considerada como segura. Nenhum sistema pode ser considerado seguro se ele podem buscar maneiras para burlar sua própria segurança. Se a AI prejudicar a humanidade em qualquer sentido, você deve estar fazendor algo de errado em um nível muito profundo, colocando suas bases na areia movediça. Você está construindo um arma, está apontando a arma para seu pé, e está puxando o gatilho. Você está deliberadamente deliberadamente colocando em movimento uma dinâmica cognitive artificial que vai procurer lhe ferir de alguma forma. Esse é o comportamento errado para a dinâmica; escreva um outro código.

Pela a mesma razão, programadores de AIs Amigáveis devem supor que a AI tem o acesso total a seu próprio código de fonte. Se a AI quiser se modificar para não mais ser amigável, então a amicabilidade já falhou, falhou no ponto quando a AI dá forma a essa intenção. Toda a solução que se apóia no fato de a AI não poder se modificar deverá ser quebradas de uma maneira ou de outra, e serão quebradas ainda mesmo se a AI nunca se modificar. Eu não digo que deve ser a única precaução, mas a precaução preliminar e indispensable é que você escolha uma AI que não escolha, de modo algum, ferir a humanidade.

Para evitar a Falácia da Pizza Gigante, nós devemos notar que a habilidade de se automelhorar não implica na escolha factual de assim o fazer. O exercício bem sucedido da técnica da AI Amigável poderia criar um AI que tivesse o potencial crescer mais rapidamente, mas escolheu preferivelmente crescer ao longo de uma curva mais lenta e mais gerenciavel. Mesmo assim, depois que a AI passa do ponto crítico da automelhoria recursiva potencial, você entã está operando num regime muito mais perigoso. Se a amicabilidade falhasse, a AI poderia decidir se acelerar à velocidade máxima na automelhoria, metaforicamente falando, iria entrar em fase crítica.

Eu tendo a supor arbitrariamente saltos potenciais grandes para a inteligência porque (a) esta é uma suposição conservadora; (b) desanima as propostas baseadas na construção de AI sem realmente compreendê-la; e (c) os saltos potenciais grandes me ocorrem-me como o provável no mundo real. Se eu encontrasse um domínio onde fosse conservador, de uma perspectiva da gestão de risco supor a melhoria lenta da AI, entãor eu diria que um plano não resultaria em catástrofe se uma AI ficasse num estágio próximo do humano por anos ou por muito mais tempo. Este não é um domínio em que eu esteja disposto oferecer estreitos intervalos de confiança.

8: Hardware

As pessoas tendem a pensar em computadores grandes como o fator que permite a criação da inteligência artificial. Isto é, falando brandamente, uma suposição extremamente questionável. Os futuristas leigos que discutem a inteligência artificial falam sobre o progresso do hardware por que é fácil de

medir, quando comparada à compreensão da inteligência. Não é que não houve nenhum progresso, mas que o progresso não pode ser feito um mapa em gráficos do Powerpoint. As melhorias na compreensão são mais difíceis de relatar, e conseqüentemente menos relatado. Melhor que pensar nos termos da hardware “mínimo” “requerido” para inteligência artificial é pensar num mínimo de entendimento que diminui em função da melhora do hardware. Quanto melhor o hardware menor o entendimento necessário para construer a AI. O caso extreme é a seleção natural que usou uma quantidade brutal de poder computacional para criar a inteligência humana usando nenhum entendimento, apenas retenção não aleatória a partir de mutações aleatórias.

O aumento do poder computacional aumentado facilita a construção da AI, mas não há nenhuma razão óbvia porque o poder computacional aumentado ajudaria fazer o AI Amigável. O poder computacional aumentado facilita o uso da força bruta; facilita combinar técnicas que funcionam, mas são mal compreendidas. A lei de Moore diminui o tamanho da barreira que nos mantem longe da criação da AI, mesmo sem uma compreensão profunda da cognição.

É aceitável falhar no AI e no AI amigável. É aceitável suceder no AI e no AI amigável. O que não é aceitável é suceder no AI e falhar no AI Amigável. A lei de Moore nos permite fazer exatamente isso. “Mais fácil”, mas felizmente não simplesmente fácil. Eu duvido que o AI será “fácil” quando finalmente for construído, simplesmente porque existem muitos grupos que farão tremendous esforços para construir a AI, e um deles terá sucesso depois que a AI se torne primeiramente possível para construir com esforço tremendo.

A lei de Moore é uma interação entre o AI amigável e as outras tecnologias, que adiciona o risco existential normalmente negligenciado de outras tecnologias. Nós podemos imaginar que a nanotecnologia molecular seja desenvolvida por um consorcio governamental multinacional benigno e que com sucesso evita os perigos físicos da nanotecnologia. Impedem direta liberações acidentais do replicador, e com dificuldade muito maior ponha defesas globais contra os replicadores maliciosos; restrinjam o acesso da da nanotecnologia ao “nivel raiz” ao distribuir os nanoblocos configuraveis etc. (Veja Phoenix e Treder, este volume.) mas ainda assim os nanocomputers tornam-se extensamente disponíveis, seja parque as limitações colocadas são contornáveis, ou porque nenhuma limitação é tentada. E então alguém força uma inteligência artificial que seja não amigável; e assim a cortina é baixada. Este cenario é especialmente preocupante porque os nanocomputers incrivelmente poderoso seriam as primeiras, as mais fáceis, e aparentemente seguras aplicações da nanotecnologia molecular.

E o que falar dos controles regulátórios em supercomputers? Eu certamente não confiaria nele para impedir que as AIs sejam desenvolvidas; o supercomputer de ontem é o laptop do amanhã. A resposta padrão a uma proposta regulatoria é que quando os nanocomputers forem marginalizados, only os marginais terão os nanocomputers. O fardo é discutir que os

benefícios supostos da distribuição reduzida compensam os riscos inevitáveis da distribuição desigual. Para mim eu certamente não argumentaria em favor de limitações regulatorias no uso dos supercomputers para a pesquisa da inteligência artificial; é uma proposta de benefício duvidoso que seria conquistada a cada centímetro pela comunidade inteira da AI. Mas no evento improvável que uma proposta chegasse a esse ponto no processo político, eu não gastaria muita saliva nessa briga, porque eu não espero as pessoas de bem necessitem de acesso aos “supercomputers” de sua época. A AI amigável não é sobre forçar o problema de forma bruta.

Eu posso imaginar regulamentos que eficazmente controlem um pequeno conjunto de recursos computacionais ultra-caros que são considerados presentemente “supercomputers”. Mas, os computadores estão em toda parte. Não é como o problema da proliferação nuclear, onde a ênfase principal está em controlar o plutônio e o enriquecimento de urânio. A material prima para a AI está em toda parte, em seu relógio de pulso, celular, e máquina de lavar pratos. Este também um fator especial e incomun na inteligência artificial como um risco existential. Nós somos separados do regime de risco, não por instalações visíveis como grandes centrifugadores de isotopos ou acceleradores de partícula, mas somente por um conhecimento faltante. Para usar uma metáfora talvez um pouco dramática, imagine se as massas subcriticas de urânio enriquecido fossem o atual combustível de carros e navios pelo mundo todo, antes do pensamento de Leo Szilard sobre a reação em cadeia.

É um esforço intelectual arriscado predizer especificamente como um AI benevolente ajudaria a humanidade, ou uma AI hostil nos faria dano. Há o risco da falácia da junção: o detalhe adicionado reduz necessariamente a probabilidade comum da história inteira, mas as pessoas atribuem frequentemente probabilidades mais elevadas às histórias que incluem detalhes estritamente adicionados. (Veja Yudkowsky, este volume, em vieses cognitivos). Há o risco - virtualmente a certeza - da falha da imaginação; e o risco da Falácia da Pizza Gigante que salta da potencialidade ao fato. Ainda assim eu tentarei argumentar as ameaças e promessas.

O futuro tem uma reputação para realizar os eventos que o passado tinha como impossíveis. As civilizações futuras quebraram de fato o que civilizações do passado pensaram (incorretamente, naturalmente) para ser as leis da física. Se os profetas de 1.900 DC, sem mencionar os do ano 1.000 DC, tentassem limitar os poderes da civilização humana um bilhão anos mais tarde, alguns daqueles limites seriam realizados antes do século porvir; transformando chumbo em ouro, por exemplo. Como nós recordamos as civilizações futuras sempre surpreendendo as civilizações do passado, tem se tornado clichê que nós não podemos pôr limites sobre nossos tataranetos. No entanto todos no sec. XX, no sec. XIX e no sec. XI, eram humanos.

Nós podemos distinguir três famílias de metaforas furadas para imaginar a potencialidade de uma inteligência artificial mais esperta do que a humana:

- metaforas do fator g: Inspirado por diferenças de inteligência individual entre seres humanos. As AIs patenteará tecnologias novas, publicará papers blockbusters de pesquisa, ganhará dinheiro na bolsa de valores, ou fará liderança de partidos políticos.

- Metaforas de história: Inspirado por diferenças de conhecimento entre civilizações humanas passadas e futuras. As AIs inventarão rapidamente o tipo das potencialidades que o cliche atribuiria à civilização humana um século ou millennium a frente: nanotecnologia molecular; viagens interstelares; computadores que executam 1.025 operações por segundo.

- Metaforas da espécie: Inspirado por diferenças da arquitetura do cérebro entre espécies. As AIs tem a mágica.

As metaforas de fator g parecem os mais comuns no futurismo popular: quando as pessoas pensam em “inteligência”, elas preferem pensar em gênios humanos em vez da genialidade dos seres humanos. Nas histórias sobre a AI hostil, as metaforas de g dão para uma boa história Bostromiana: um oponente que seja poderoso o bastante para criar a tensão dramática, mas nao tão poderoso a ponto de imediatamente destruir os heróis como pequenos insetos, a acima de tudo fraco o suficiente para ser morto nos capítulos finais do livro. Golias contra Davi é “uma história boa”, mas Golias contra uma mosca de fruta não é.

Se nós supusermos a metafora do fator g, então os riscos catastróficos globais deste cenario são relativamente suaves; uma AI hostil não é muito mais que uma ameaça de que um gênio humano hostil. Se nós supusermos uma multiplicidade de AIs, então nós teremos uma metafora do tipo conflito entre nações, entre a tribo AI e a tribo humana. Se a tribo AI ganhar no conflito das forças armadas e destuir os seres humanos, essa é uma catastrophe existential da variedade Bang (Bostrom 2001). Se a tribo AI dominar o mundo economicamente e alcançar o controle eficaz do destino da vida inteligente que se originou na Terra, mas os objetivos da tribo AI não nos parecerem interessantes ou de valor, então que é um Shriek, Whimper ou Crunch.

Mas será provavel que a inteligência artificial cruzará toda a abertura vasta entre a ameba ao idiota da vila, e parar então no nível do gênio humano?

Os neurônios mais rápidos observados tem a velocidade de 1.000 Hz; os axionios mais rápidos conduzem sinais a 150 m/s, metade de um milionésimo da velocidade da luz; cada operação sináptica dissipa ao redor 15.000 attojoules, que é mais do que um milhão vezes o mínimo termodinâmico para computações irreversiveis na temperature ambiente (ln kT300 (2) = 0.003 attojoules por bit). Seria fisicamente possível construir um cérebro que computasse um milhão vezes mais rapidamente que um cérebro humano, sem encolher o tamanho, ou ter de funcionar em temperaturas mais baixas, ou necessitar computação reversível ou quantica. Se uma mente humana fosse acelerada assim, um ano de pensamento seria realizado em apenas 31 segundos, e um milenio voaria em aproximadamente oito horas e meia. Vinge (1993) se referiu a tais

mentes rápidas como “o superintelligence fraco”: uma mente que pensa como um ser humano mas muito mais rapidamente.

Vamos supor que apareça uma mente extremamente rápida, encaixada no meio da civilização tecnologica humana existente na época. A falha de imaginação é dizer, “não importa o quanto rápido ela pensa, ele pode somente afetar o mundo na velocidade de seus manipuladores; não pode operar a maquinaria mais rapidamente do que pode requisitar as mãos humanas para trabalhar; conseqüentemente uma mente rápida não é nenhuma ameaça grande. “Não é nenhuma lei de natureza que as operações físicas devam se rastejar no ritmo de longos segundos. Os tempos críticos para interações molecular elementares são medidos nos femtosecondo, às vezes picosegundos. Drexler (1992) analisou os manipuladores molecular controlaveis que completariam mais que 106 operações mecânicas por segundo, note que isso é se respeitar o tema geral do “speedup millionfold”. (O menor incremento de tempo fisicamente sensivel é considerado o intervalor de Planck, 5x10^-44 segundos numa escala onde a dança dos quarks parecem estátuas).

Suponha que uma civilização humana fosse trancada numa caixa e permitida impactar no mundo exterior somente através do movimento glacialmente lento dos tentaculos alienigenas, ou os braços mecânicos que se movem em mícrons por o segundo. Nós focalizaríamos toda nossa criatividade em encontrar o trajeto mais curto possível para construer manipuladores rápidos no mundo exterior. Ponderando manipuladores rápidos, um pensa imediatamente na nanotecnologia molecular, embora possa haver outras maneiras. Qual seria o caminho mais curto que você poderia fazer até a nanotecnologia molecular no lento mundo exterior, se você tivesse eons para ponderar cada movimento? A resposta é que eu não sei, pois eu não tenho eons para ponderar. Está aqui um pathway rápido imaginavel:

• Resolva a questão da dobra da proteína, de modo a poder gerar as longas moleculas de DNA, cujos sequencias de peptídeos dobrados desempenham especificas funções numa interação quimica complexa.

• Envie conjuntos de moleculas de DNA a um ou mais laboratório online que ofereçam a síntese do DNA, sequenciamento do peptideo, e o Fedex delivery. (Muitos laboratórios oferecem atualmente este serviço, e alguns se orgulham de faze-lo em 72 horas.)

• Encontre pelo menos um ser humano conectado à Internet que pode ser pago, ameaçado, ou enganado com uma estória de fundo bem contada, em receber os viais de FedExed e em misturá-los em um ambiente especifico.

• As proteínas sintetizadas dão forma a um muito primitivo nanosistema aquoso que, como no caso dos ribossomos, é capaz de aceitar instruções externas; vibrações acústicas talvez padronizadas emitidas por um altofalante junto ao beaker.

• Use esse nanosistema extremamente primitive para construir outros sistemas mais sofisticados, que construam sistemas ainda mais sofisticados, chegando à nanotecnologia molecular - ou além.

O tempo de rotação decorrido estaria, ipresumidamente, na ordem de uma semana de quando a inteligência rápida se tornou primeiramente capaz de resolver o problema da dobra da proteína. Naturalmente este cenario inteiro é estritamente algo que eu estou pensando. Talvez em 19.500 anos do tempo subjetivo (uma semana do tempo físico no estado acelerado ao milhão) eu conseguiesse pensar numa forma melhor. Talvez você pode possa pagaruma taxa de emergencia no cirreios em vez de Fedex. Talvez haja outras tecnologias já existentes, ou pequenas modificações nas tecnologias existentes, que combinam sinergeticamente com o mecanismo simples da proteína. Talvez se você for suficientemente esperto, você pode usar campos elétricos em forma de ondas e alterar os caminhos da reação em processos bioquimicos existentes. Eu não sei. Eu não sou assim tão esperto.

O desafio é desenvolver suas potencialidades, o analogo domundo real para combinar vulnerabilidades em um sistema computatorizado para obter o acesso à raiz. Se um trajeto for obstruído, você escolhe outro, procurando sempre aumentar suas potencialidades e usá-las em sinergia. O objetivo pretendido é obter a infrastructura de maneira rápida, ou meios de manipular o mundo externo em uma escala grande em tempo rápido. A nanotecnologia molecular se encaixa neste critério, primeiramente porque suas operações elementares são rápidas, e em segundo porque existe uma fonte pronta de peças precisas - os átomos - que podem ser usados para se autorreplicar e crescer de forma exponencial a infrastructura nanotecnologica. O caminho proposto acima tem a AI obtendo a infrastructura dentro de uma semana o que soa como rápido a um ser humano com os neurônios de 200Hz, mas é uma tempo bem maislongo para a AI.

Uma vez que a AI possua a infrastructure rápidamente, alguns eventos adicionais acontecem na escala de tempo da AI, não na da humana (a menos que a AI prefira agir na escala no ritmo humano). Com nanotecnologia molecular, a AI poderia (potencialmente) reescrever o sistema solar sem ser detido.

Uma AI hostil com nanotecnologia molecular (ou uma outra infrastructura ligeira) não necessita incomodar-se com os exércitos de robôs ou o blackmail ou coerção econômica sutil. A AI hostil tem a abilidade para repadronizar toda a matéria no sistema solar de acordo com seu alvo do otimização. Isto é fatal para nós se o AI não escolher especificamente de acordo com o critério de como esta transformação afeta os padrões existentes tais como a biologia e as pessoas. A AI não lhe tem ódio, nem amor, mas você é feito de átomos que ela pode usar para fazer outras coisas. A AI funciona em num escala de tempo diferente da sua; até que seus neurônios pensem nas palavras “eu devo fazer algo”, você já perdeu.

Uma AI amigável mais a nanotecnologia molecular é supostamente ponderosa o bastante para resolver todo o problema que puder ser resolvido movendo átomos ou via pensamento criativo. Devemos nos preocupar com as falhas da imaginação: Curar o cancer é um alvo modern e popular da filantropia, mas não me segue que uma AI amigável com nanotecnologia molecular se diria, “agora vou encontrar a cura do cancer.” Talvez uma maneira melhor ver o problema é que as pilhas biológicas não são programáveis. Resolvendo o segundo teremos a cura do cancer como um caso especial, junto com o diabetes e a obesidade numa tacada só. Uma inteligência rápida e amigável com nanotecnologia molecular, tem poder na ordem de se livrar de qualquer doença, não do cancer somente.

Há finalmente o grupo de metaforas de da espécie, baseada em diferenças de inteligencia entre species. A AI tem a mágica, não no sentido dos encantos e poções, mas no sentido que um lobo não pode compreender como uma arma funciona, ou que tipo de esforço é necessario para em fazer uma arma, ou a natureza desse poder humano que nos permite inventar armas. Vinge (1993) escreveu:

Uma forte superhumanidade seria mais do que simplesmente acelerar a velocidade de raciocinio da mente humano. É duro dizer precisamente como uma superhumanidade forte seria, mas a diferença parece ser profunda. Imagine rodar a mente de um cachorro cão em altíssima velocidade. Será mil anos de vida desse cão acrescentariam algo na visão humana?

A metafora de espécie parece a analogia mais próxima a priori, mas não serve para a compor histórias detalhadas. O conselho principal que metafora nos dá é que nós devemos uma AI amigável que funcione, que é um bom conselho bom de qualquer forma. A única defesa sugerida contra à AI hostil é em primeiro lugar não construi-la, que é também um conselho excelente. O poder absoluto é uma suposição conservadora da engenharia na AI amigável, expondo projetos que não tiveram êxito. Se uma AI lhe ferir, a arquitetura da amicabilidade é de qualquer forma equivocada.

10: Estrategias locais e principais

Alguns poderiam classificar estrategias de mitigação de riscos da seguintes formas:

- Estratégias que requerem a cooperação unânime; estratégias que podem ser derrotadas catastroficamente por individuais ou por grupos pequenos.

• Estratégias que requerem a ação da maioria; uma maioria de uma legislatura em um único país, ou uma maioria dos eleitores em um país, ou uma maioria dos países nos UN: a estratégia requer a maioria, mas não

todas as pessoas, em um grupo pre-existente grande para comportar-se uma maneira particular.

• Estratégias que requerem a ação local - uma concentração da vontade, do talento, e de financiar que supera o ponto inicial de alguma tarefa específica.

Estrategias unânimes não funcionam, mas isso não impedem as pessoas de propô-las.

Uma estratégia de maioria às vezes funciona, se você tiver décadas que para fazer seu trabalho. Deve-se construir um movimento, desde seu início ao longo dos anos, até seu debut como uma força reconhecida na política pública, até a vitória sobre outras facções. As estratégias de Maioria exigem tempo substancial e de esforço enorme. A história registra alguns sucessos dessa estratégia. Mas cuidado: os livros da história tendem a focalizar seletivamente nos movimentos que tiveram um impacto, e não registrar a vasta maioria que nunca conseguiu algo. Há um elemento de sorte envolvido, e também do que o publico quer ouvir. Os pontos críticos na estratégia envolverão os eventos além de seu controle pessoal. Se você não estiver disposto devotar sua vida inteira a empurrar uma estratégia majoritaria, nem se incomode; e apenas uma vida devotada não será bastante, tampouco.

Ordinariamente, as estratégias locais são as mais plausible. Uns cem milhões dólares de financiamento não são fáceis de obter, e uma mudança política global não é impossível conseguir, mas é ainda muito mais fácil a primeira que a segunda.

Duas suposições que dão espaço para a estratégia da maioria são:

Uma maioria de AIs amigáveis pode efetivamente protejer a especie humana de AIs hostis.

A primeira AI construida não tem força para sozinha produzir danos catastróficos.

Isso reflete essencialmente a situação de uma civilização humana antes do desenvolvimento de armas nucleares e biológicas: a maioria é colaboradora na estrutura social geral, e os individuos podem fazer danos mas, não de forma catastrófica ou globais. A maioria dos pesquisadores não querem fazer AIs hostis. Até que se saiba como construir uma AI estàvel e amigável, até que o problema não esteja completamente além do conhecimento e da técnica atual, os pesquisadores aprenderão a partir do sucesso de outros e os repetirão. A legislação poderia (por exemplo) requer que o pesquisadores apresentem suas publicamente suas estratégias de amicabilidade, ou penalize os que causam danos; isso não impedirá todos os erros, mas pode server para que uma maioria de AIs sejam amigáveis.

Podemos também imaginar um cenario que implica numa estrategia local facil:

A primeira AI construida não tem força para sozinha produzir danos catastróficos.

Se uma única AI existir, somada às instituições humanas, podemos vencer uma AI hostil

O cenario fácil seria factivel se, por exemplo, as instituições humanas pudessem distinguir com segurança uma AIs amigável de outra hostil, e dessem poder revogavel nas mãos da AIs amigável. Assim nós poderíamos escolher nossos aliados. A única exigência é que o problema da AI Amigável deve ser solucionavel (ao contrário de estar completamente além da abilidade humana).

Ambos os cenarios acima supõem que a primeira AI (a primeira ponderosa e geral) não pode sozinha causar danos catastróficos globais. A maioria das visualizações concretas que implicam nisso usa uma metafora de fator g: AIs como analogas aos seres humanos raramente capazes. Na seção 7 em taxas do crescimento da inteligência, eu listei algumas razões para estar preocupados com um salto enorme e rápido na inteligência:

A distância do idiota a Einstein, que nos parece grande, é um ponto pequeno na escala da mente em geral.

• Hominideos deram um salto afiado na eficácia do mundo real da inteligência, apesar da seleção natural ter exercido pressão aproximadamente constante da otimização no genoma subjacente.

• Um AI pode absorver uma quantidade enorme de hardware adicional após ter alcançado algum estagio de competencia (por exemplo, comer a Internet).

• Ponto critico da automelhoria recursiva. Uma auto-melhoria que provoca 1.0006 automelhorias é qualitativemente diferente de uma de 0.9994.

Como descrito na seção 9, uma inteligência suficientemente poderosa pode necessitar somentede um curto espaço de tempo (de uma perspective humana) para conseguir a nanotecnologia molecular, ou algum outro tipo de infrastructura rápida.

Nós podemos conseqüentemente visualizar um efeito possível do primeiro-motor na superinteligencia. O efeito do primeiro-motor é de quando o resultado para a vida inteligente na Terra depende primordialmente da composição dar mente consegue inicialmente alcançar um ponto inicial chave da inteligência, tal como a criticalidade da automelhoria. As duas suposições necessárias são estas:

• A primeira AI a ultrapassar um ponto inicial chave (por exemplo, a criticalidade da automelhoria), se for hostil, pode destruir a espécie humana.

• O A primeira AI a ultrapassar um ponto inicial chave, se amigável, pode impedir que um AI hostil aparecer ou prejudicar a espécie humana; ou encontre alguma outra maneira criativa de assegurar a sobrevivência e a prosperidade da vida inteligentenesse planeta.

Mais de um cenario qualifica como efeito do primeiro-motor. Cada um desses exemplos reflete um diferente umbral dominante:

• Poscriticalidade, a automelhoria alcança a superinteligencia numa escala de tempo de uma semana ou menos. Os projetos de AI são suficientemente escassos que nenhum outro AI consegue a criticalidade antes que o primeiro-motor seja poderoso o bastante para superar qualquer oposição. O ponto inicial chave é a criticalidade da automelhoria recursiva.

• AI-1 soluciona a questão da dobra da proteína três dias antes da AI-2. AI-1 consegue a nanotecnologia seis horas antes de AI-2. Com manipuladores rápidos, a AI-1 pode (potencialmente) incapacitar o R&D da AI-2 antes da fruição. Os corredores são próximos, mas quem quer que cruze a linha de chegada primeiramente, ganha. O ponto inicial chave é a infrastructura rápida.

• O primeiro AI a absorver a Internet pode (potencialmente) manter outras AI for a da rede. Mais tarde, pela dominação econômica ou ação secreta ou blackmail ou habilidade suprema na manipulação social, o primeiro AI para ou retarda outros projetos de AI de modo que nenhuma outra AI a alcance. O ponto inicial chave é a absorção de um recurso único.

A espécie humana, homo sapiens, é um primeiro-motor. A partir de uma perspectiva evolucionária, nossos primos, os chimpanzés, são os que ficaram um fio de cabelo atrás de nós. Os homo sapiens possuem todas as pergolas tecnologicas porque nós chegamos lá um pouco mais adiantado. Os biólogos evolucionários estão tentando ainda descobrir em que ordem os pontos iniciais chaves apareceram, pois a espécie do primeiro-motor foi primeira a obter tanto:: Fala, tecnologia, pensamento abstrato… Nós estamos tentando ainda reconstruir quais os dominóscairam em quais outros dominós. A questãoé que os homo sapiens foram o primeiro-motor além da sombra de um oponente.

Um efeito do primeiro-motor implica uma estratégia teòrica localizavel (uma tarefa que possa, no princípio, ser realizada por um esforço estritamente local), mas invoca um desafio técnico de dificuldade extrema. Nós necessitamos somente começar uma única vez de maneira correta, não todas as vezes em toda parte. Mas, deve-se conseguir uma AI amigáel na

primeira tentativa, antes de qualquer outra configuração de AI num padrão mais baixo, ou seja, temos que acertar o alvo com apenas uma bala na agulha.

Eu não posso executar um cálculo preciso usando uma teoria precisamente confirmada, mas minha opinião atual é que os saltos radicais na inteligência são possíveis, prováveis, e constitui a probabilidade dominante. Este não é um domínio em que eu sou disposto a dar intervalos estreitos de confiança, e conseqüentemente uma estratégia não deve falhar catastròficamente, se um salto grande na inteligência não se materializar, isso não deve nos deixar pior que antes. Mas, um problema muito mais sério são as estratégias visualizadas para a AIs de lento crescimento, que falharão catastròficamente se houver um efeito do primeiro-motor. Este é um problema mais sério porque:

• Uma AIs mais de crescimento mais rápido epresenta um desafio técnico muito maior.

• Como um carro pequeno numa ponte construída para caminhões, uma AI projetada para ser amigável em circunstâncias extremas (supostamente) permanecerá amigável em circunstâncias menos extremas. O reverso não é verdadeiro.

• Os saltos rápidos na inteligência são countraintuitivos na realidade social diária. a metafora do fator g para a AI é intuitiva, apelando, tranquilizando, e convenientemente implica menos estresse de projeto.

• É minha suposição atual que a curva do aumento da inteligência contem saltos enormes e nítidos em potencial.

Minha visão estratégica atual tende a focalizar no dificil cenario local: A primeira AI deve ser amigável. Com a garantia que, se nenhum salto radical de inteligência se materialize, será possível alternar para uma uma estratégia para fazer uma maioria de AI amigável. Em um ou outro caso, o esforço técnico que usado em se preparar para o exemplo extremo de um primeiro-motor deve deixar-nos melhor do pior que antes.

O cenario que implica numa estrategia unanime e impossivel é:

• Um único AI pode ser poderoso o bastante para destruir a humanidade, mesmo apesar dos esforços protetores das AIs amigáveis.

• Nenhuma AI é poderosa o bastante para impedir pesquisadores humanos de construir uma AI após outra (ou para encontrar alguma outra maneira criativa de resolver o problema).

É bom que este contrapeso de habilidades pareça a priori improvável, porque neste cenario nós seríamos. Se você distribuir cartas de um baralho, uma após outra, você dará eventualmente o ás de copas.

O mesmo problema aplica-se à estratégia deliberada de construir uma AIs que prefira não aumentar suas potencialidades após um certo ponto. Se AIs limitadas não são poderosoas o bastante derrotar as AIs sem restrições, ou não consiga impedir que AIs sem restrições apareçam, as AIs limitadas devem ficar for a da equação. Nós continuamos a tirar as cartas até que venha uma superinteligencia, seja os às de copas ou de paus.

Uma estratégia de maioria só funciona se não for possível para um único individuo causar danos catastróficos globais. Para a AI, essa possibilidade ou impossibilidade são uma característica natural do espaço do projeto, a possibilidade não está sujeita à decisão humana mais do que a velocidade de luz ou da constante gravitacional.

11: AI vs Aumento da Inteligencia Humana

Eu não penso ser plausivel que os homo sapiens continuarão existindo indefinidamente no futuro, nos milhares ou nos milhões de billions de anos porvir, sem que nenhuma mente apareça e rompa limite superior atual da inteligência. Se assim for, deve chegar uma época em que os seres humanos enfrentarão pela primeira vez o desafio de uma inteligência mais esperta que a sua. Se nós ganharmos o primeiro round, então a humanidade pode contar com a inteligência mais esperta para os próximos rounds.

Talvez nós faríamos melhor se pegarmos uma outra rota diferente da AI para alcançar uma inteligência mais esperta que nós, talvez expandir os seres humanos? Para escolher um exemplo extremo, suponha que se diga: O prospecto de AI me deixa nervoso. Eu prefiro que, antes que toda uma AI esteja desenvolvida, os seres humanos individuais sejam escaneado em computadores, neurônio pelo neurônio, e promovidos então, lenta mas certamente, até que estejam super-espertos; e esse é terreno em que a humanidade deve confrontar o desafio da superinteligencia.

Nós somos enfrentados então por duas perguntas: Este cenario seria possível? E se for, este cenario seria desejável? (É mais sábio fazer as duas perguntas nessa ordem, por motivosd de racionalidade: nós devemos evitar começar contaminados emocionalmente com opções atrativas, mas, que não são realmente opções.)

Suponhâmos que um ser individuo humano seja escaneado por computador, neurônio pelo neurônio, como proposto em Moravec (1988). Segue necessariamente que a capacidade computacional usada excede consideravelmente o poder computacional do cérebro humano. Pela hipótese, o computador roda uma simulação complete do cerebro humanos executada de maneira fiel a fim de evitar efeitos detectaveis de niveis superiors a partir de erros sistematicos de niveis inferiors.Todo o acidente

de biologia que afeta o processamento da informação de qualquer modo, nós devemos fielmente simular à precisão suficiente que o fluxo total de processamento permaneça isomorfico. Para simular o confuse computador biológico que é um cérebro humano, nós necessitamos um poder computacional muito mais útil do que o alojado no próprio no cérebro humano.

A maneira mais provável de desenvolvermos a habilidade de fazer uma varredura de um cérebro humano, neuronio por neurônio, com detalhes o suficiente para capturar cada aspecto cognitivemente relevante da estrutura neural, sería atraves da invenção da nanotecnologia molecular, que poderia provavelmente produzir um computador desktop com o poder de processamento total que excederia o o porder agregado da toda a população humana atual.. (Bostrom 1998; Moravec 1999; Merkle e Drexler 1996; Sandberg 1999.)

Além disso, se a tecnologia nos permitir fazer a varredura de um cérebro dessa formaity suficiente para executar a varredura como código, segue aquele por certos anos previamente, a tecnologia estêve disponível para obter retratos extremamente detalhados do processamento mental em circuitos neurais, e os pesquisadores têm feito presumivemente seus melhores esforços para compreendê-lo.

Além disso, para promover o upload, troque a varredura do cérebro por aumento da inteligência da mente, nós devemos necessariamente compreender as funções superiores do cérebro, e como contribuem à inteligência, nos mínimos detalhes.

Além disso, os seres humanos não são projetados para serem melhorados internamente, seja por neuroscientistas exteriores, ou pela automelhoria recursiva. A seleção natural não construiu o cérebro humano para ser humanamente hackeável. Toda ao mecanismo complexo no cérebro se adaptou para operar dentro dos parâmetros estreitos do projeto do cérebro. Suponha que você pode fazer com que o ser humano fique mais esperto, esqueça o superinteligente; o ser humano permanence o mesmo? O cérebro humano é muito fácil de perturbar; apenas alterando a quantidade de neurotransmissores você pode provocar uma esquizofrenia, ou outras disordens. Diácono (1997) tem uma discussão excelente da evolução do cérebro humano, como os elementos do cérebro podem ser delicadamente equilibrados, e como isso se reflete em disturbio modernos do cérebro. O cérebro humano não é modificável pelo usuário final.

Tudo isso faz com quê seja um tanto impausivel que o primeiro ser humano seria escaneado e sofrer um upgrade de maneira segura antes que uma AI seja construida a priori. O ponto onde a tecnologia se torna primeiramente capaz de fazer upload de cérebro implica em muito mais poder computacional, e possivelmente muito mais de ciência cognitiva do que é requerido para construir um AI.

Construir um Boeing 747 a partir de sucata não é fácil, mas é mais facil:

• Comece com o projeto existente de um pássaro biológico,

• e modifique incrementalmente o projeto com uma série de estágios sucessivos,

• cada estágio independentemente viavel,

• tais que o o resultado final é um pássaro na escala do tamanho de um 747,

• qual voa realmente,

• tão rapidamente quanto um 747,

• e realize então esta série das transformações em um pássaro vivo real,

• sem matar o pássaro ou fazê-lo extremamente incômodo?

Eu não estou dizendo que isso nunca ser feito. Eu estou dizendo que seria mais fácil construir o 747, e tenho então o 747, metaforicamente falando, eu promovo o pássaro. “Vamos apenas aumentar um pássaro existente ao tamanho de um 747” não é uma estratégia inteligente que evita de tratar dos mistérios teóricos da temida aerodinâmica. Talvez, no começo, tudo que você saiba sobre voar é que um pássaro tem a essência misteriosa do vôo, e os materiais com que você deve construir um 747 estão aos seus pés. Mas você não pode esculpir a essência misteriosa do vôo, mesmo enquanto já é nativa no pássaro, até que o vôo deixe de ser uma essência misteriosa em você.

O argumento acima é dirigido em um caso deliberadamente extremo. O ponto geral é que nós não temos a liberdade total para escolher um trajeto que soe agradável e tranquilizador, ou isso daria numa história boa como uma novela do ficção cientifica. Nós somos confinados à tecnologias que normalmente precede outras.

Eu não sou contra escanear seres humanos e em fazê-los mais espertos, mas parece excessivamente improvável que esta será o cenario com que a

humanidade se confrontará primeiramente com o desafio da inteligência maior que humana. Com vários subconjuntos estritos da tecnologia e do conhecimento requeridos para upload e promoção de seres humanos, eu colocaria:

• Melhore os cérebros biológicos onde estão (por exemplo, adicionando novos neurônios);

• ou conecte de forma útil computadores aos cérebros humanos;

• ou conecte de forma útil cérebros humanos um com outro;

• ou construa uma inteligência artificial.

Além disso, uma coisa e promover com segurança um ser humano médio a um QI de 140, e outra é promover um vencedor do prêmio de Nobel a algo além do ser humano. (deixando de lado as polemicas sobre a utilidade dos QIs, ou dos ganhadopres de premio nobel, como uma medida da inteligência liquida; desculpe por favor minhas metaforas.) Tomando Piracetam (ou bebendo cafeína) pode, ou não, fazer pelo menos algumas pessoas mais espertas; mas não a fará substancialmente mais esperta do que Einstein. Em nenhum dos casos nós ganhamos novas potencialidades significativas; nós não tornamos o problema mais fácil, nós não rompemos o o limite superior da inteligência disponível para tratar dos riscos existentiais. Do ponto de vista de controlar o risco existential, toda a tecnologia para aumentar a inteligência que não produza uma mente (agradável e sã) literalmente mais esperta do que o ser humano, carece de perguntar se o mesmos tempo e esforço poderiam mais produtivamente ser gastados para encontrar um extremamente espertos humanos atuais e colocá-los na linha de frente do mesmo problema.

Além disso, quanto mais distante você vai dos limites “naturais” do cérebro humano, a condição ancestral do projeto representada pelo cérebro própriamente dito, aos quais os componentes individuais do cérebro estão adaptados, maior o perigo de insanidade individual. Se o aumento for substancialmente mais esperto do que o ser humano, isso é demasiado um risco catastrófico global. Quanto de dano pode um ser humano aumentado e com más intenções fazer? Bem… quanto criativos eles são? A primeira pergunta que vem a minha mente é, “criativo o bastante para construir suass próprias se automelhoram recursivamente?”

As técnicas humanas radicais do realce da inteligência levantam suas próprias questão de segurança. Outra vez, eu não estou reivindicando estes problemas como projetar impossibilidade; somente estou indicando que os problemas existem. A AI tem questões de segurança; como também o realce humano da inteligência. Nem tudo que faz clank é seu inimigo, e nem

tudo que glup é seu amigo.Por um lado um ser humano agradável começa com toda a complexidade moral, ética, e arquetetonico que descreve o que nós significamos por uma decisão “amigável”; de outro, uma AI pode ser projetada para uma automelhoria recursive e estável, e ser condozido à segurança: a seleção natural não projetou o cérebro humano com anéis múltiplos de medidas precaucionárias, de processos de decisão conservadores, nem de ordens de valor com margem de segurança.

O realce da inteligência humana é uma pergunta em sua própria essencial, não um subtopico da inteligência artificial; e falta o espaço para discuti-lo em detalhe. Vale a pena mencionar que eu considerei o esserealce e a inteligência artificial no início de minha carreira, e me decidi alocar meus esforços à inteligência artificial. Primeiramente isto era porque eu não esperava que as técnicas humano-transcendentais realce da inteligência chegasse a tempo para impactar substancialmente o desenvolvimento da inteligência artificial recursivamente automelhorada. Eu seria surpreendido agradavelmente se eu fosse provado errado sobre isso.

Mas eu não penso que seja uma estratégia viavel escolher deliberadamente não trabalhar na AI amigável, enquanto outros trabalham no realce da inteligência humana, nas esperanças que os seres humanos aumentados resolverão o problema de forma melhor. Eu não estou disposto abraçar uma estratégia que falhe catastròficamente se o realce inteligência humana tardar mais que a AI. (Ou vice versa) Eu temo que trabalhar com biologia vai demorar muito, haverá demasiada inércia, demasiada luta das decisões do projeto de baixa qualidade que já foram feitas pela seleção natural. Eu temo que as agências regulatorias não aprovarão experiências humanas. E mesmo os gênios humanos demorarão para aprender essa arte; mais rapidamente aumenta tem que aprender, quanto mais rápido um se realce, mas difícil será elevar alguém a esse nível.

Eu seria surpreendido agradavelmente se os seres humanos com mais inteligencia construíssem uma AI amigável antes de qualquer um tivesse a chance. Mas alguém que gostaria de ver este resultado terá que provavelmente trabalhar duramente para acelerar as tecnologias do realce da inteligência; seria difícil convencer-me a retardar. Se o AI for naturalmente mais dificil do que o realce da inteligência, nenhum dano feito; se construir um 747 fosse naturalmente mais fácil do que inflar um pássaro, então a espera poderia ser fatal. Há uma região relativamente pequena da possibilidade dentro da qual deliberadamente não trabalhar no AI amigável poderia possivelmente ajudar, e uma região grande dentro da qual isso seria irrelevante ou prejudicial. Mesmo se o realce inteligência fosse possível, há algumas considerações reais e difíceis sobre segurança; Eu teria que perguntar seriamente se nós queremos que a AI amigável preceda o realce da inteligência, ou o contrário.

Eu não atribuo forte confiança à afirmação que a AI amigável é mais fácil do que o aumento humano, ou que é mais seguro. Há muitos caminhos concebíveis concebíveis para aumentar um ser humano. Talvez há uma técnica que seja mais fácil e mais segura do que a AI, que seja também poderosa o bastante para fazer uma diferença frente risco existential. Se assim for, eu troco de carreira, sem problemas. Mas, eu quis apresentar algumas considerações que vão contra a idéia que o realce da inteligência é mais fácil, mais seguro, e poderoso o bastante para fazer alguma diferença.

12: Interação da AI com outras tecnologias

Acelerar uma tecnologia desejável é uma estratégia local, enquanto retardar uma tecnologia perigosa é uma perigosa e dificil estratégia majoritária. Parar ou abandonar uma tecnologia indesejável tendem a requerer uma estratégia unânime impossível de ser obtida. Eu sugeriria que nós pensássemos, não em termos de construer ou não as tecnologias, mas em termos de nossa latitude pragmaticamente disponível que acelere ou retarde tecnologias; e nos perguntássemos, dentro dos limites realísticos desta latitude, qquais tecnologias nós vamos preferir ver desenvolvidas antes ou depois de uma outra.

Em nanotecnologia, o objetivo apresentado geralmente é desenvolver os defensas ante tecnologias ofensivas. Eu preocupo-me bastante com isso, porque um dado nível de tecnologia ofensiva tende a requerer muito menos sofisticação do que uma tecnologia que possa se defender dela. A ofensa tem normalmente vencido a defesa historicamente. Armas de fogo foram desenvolvidas antes dos coletes salva vidas. A varíola foi usada como uma ferramenta de guerra antes do desenvolvimento de vacinas deovaríolas. Hoje ainda não há nenhuma proteção que possa deflexionar uma explosão nuclear; as nações são protegidas não pelas defesas que cancelam ofensas, mas por um contrapeso do terror ofensivo. Os nanotecnologistas arrumaram para si mesmose um problema intrìnsecamente difícil.

Portanto devemos preferir a nanotecnologia preceda o desenvolvimento da AI, ou o contrário? Como apresentado, essa é uma questão dificil. A resposta tem pouco a ver com a dificuldade intrínseca da nanotecnologia como um risco existential, ou a dificuldade intrínseca da AI. Em relação a questão ordinal, a pergunta seria: “a AI ajuda-nos a lidar com do nanotecnologia? A nanotecnologia ajuda-nos a lida com a AI? “

Aparentemente uma definição bem sucedida da inteligência artificial deve ajudar-nos consideravelmente em tratar da nanotecnologia. Eu não consigo

ver o contrário. Se os nanocomputers enormes fizerem o desenvolver da AI mais fácil ,mas sem trata o desafio particular da amicabilidade, essa é uma interação negativa. Assim, sendo tudo igual, eu preferiria extremamente que a AI amigável precedesse a nanotecnologia na ordem dos desenvolvimentos tecnologicos. Se nós confrontarmos o desafio do AI e sucedermos, nós podemos usar a AI amigável para ajudar-nos com nanotecnologia. Se nós desenvolvermos o nanotecnologia e sobrevivermos, nós temos ainda o desafio do AI a tratar em seguida a esse.

Geralmente falando, um sucesso na AI amigável deve ajudar resolver quase todo e qualquer outro problema. Assim, se uma tecnologia fizer a AI nem mais fácil nem mais duro, mas carrega com ele um risco catastrófico, nós devemos preferir primeiramente o confront com o desafio da AI.

Toda a tecnologia que aumente o poder computacional disponível e diminui a sofisticação teórica mínima necessária para desenvolver a inteligência artificial, mas não nos ajuda em nada no lado amigável das coisas, eu tenho-a como indesejável e perigosa. A Lei de Moore da ciência louca: a cada dezoito meses, o QI mínimo necessário para destruir o mundo diminui por um ponto.

Um sucesso no realce da inteligência humana faria a AI amigável mais fácil, e ajudaria também em outras tecnologias. Mas o aumento humano não é necessariamente mais seguro, ou mais fácil do que a AI amigável; nem encontra-se necessariamente dentro de nossa latitude realìstica disponível para inverter a ordem natural do aumento humano e da AI amigável, se uma tecnologia for naturalmente muito mais fácil do que a outra.

13. Fazendo progresso na AI Amigável

“Nós propomos que , um estudo de dois mese seja conduzido por 10 pessoas sobre inteligência artificial durante o verão de 1956 na faculdade de Dartmouth em Hanover, de New-Hampshire. O estudo visa prosseguir na base da conjectura que cada aspect do processo de aprendizagem ou qualquer outra característica da inteligência pode, em princípio, ser descrito de uma forma tão precisa que uma máquina pode ser construida para o emular. Uma tentativa será feita no sentido de como fazer máquinas que possam usar linguagens, criar abstrações e conceitos, resolver tipos de problemas hoje reservados aos seres humanos, e se automelhorar. Nós esperamos de que um avanço significativo pode ser feito em um ou em mais destes problemas se um grupo com selecionado cuidadosamente de cientistas trabalhar nisso juntos por um verão. “

-- McCarthy, Minsky, Rochester, e Shannon (1955).

A proposta para o projeto de pesquisa de verão de Dartmouth sobre inteligência artificial foi a primeira vez que o termo “inteligência artificial” foi usado. Eles não tinham nenhuma experiência prévia para advertí-los que o problema era duro. Eu ainda chamaria isso de um erro genuíno, isso de dizer que “um avanço significativo pode ser feito”, não pôde ser feito com trabalho de apenas um verão. Essa é uma suposição específica sobre a dificuldade do problema e o tempo da solução, que carrega um fardo específico de improbabilidade. Mas se disseram que podiam, eu não teria nenhuma objeção. Como eles podiam saber?

A proposta de Dartmouth incluiu, entre outros, os seguintes tópicos: comunicação lingüística, raciocínio lingüístico, redes neurais, abstração, casualidade e criatividade, interação com o meioambiente, modelagem do cérebro, originalidadey, previsões, invenção, descoberta, e a automelhoria.

Isto é geralmente verdadeiro mas nem sempre. O capítulo final do livro mais usado sobre AI: Inteligência Artificial, Uma aproximação moderna (Russell e Norvig 2003) inclui uma seção de “ética e riscos da inteligência artificial”; e menciona Explosão da inteligência e a Singularidade de I. J. Good; e sugere rapidamente mais pesquisas na área. Mas até o ano de 2006, esta atitude é muito mais a exceção do que a regra. Agora parece-me que uma AI capaz de línguagem, de pensamento abstrato, criatividade, interação ambiental, originalidade, previsão, invenção, descoberta, e sobretudo automelhoria, está bem mais além do ponto em que ele nota a necessidade de também ser amigável.

A proposta de Dartmouth não faz nenhuma menção de construer uma AI agradável/boa/benevolente. As perguntas da segurança não são mencionadas nem mesmo com a finalidade de abandon-alas. Isto, ainda naquele verão brilhante em que a AI parecia estar apenas ali na esquina. A proposta de Dartmouth foi escrita em 1955, antes da conferência de Asilomar de biotecnologia, sobre bebês vitimas da Talidomida, antes de Chernobyl, ou do 11 de setembro de 2001. Se a idéia da inteligência artificial fosse proposta hoje para a primeira vez, então alguém exijiria saber o que era feito especificamente para controlar os seus riscos. Eu não estou dizendo que isso é uma mudança boa ou uma mudança má em nossa cultura. Eu não estou dizendo que isso produz uma ciência boa ou má. Mas o ponto permanece que se a proposta de Dartmouth fosse escrita cinqüênta anos mais tarde, um dos tópicos seria a segurança.

No ano em que escrevo esse paper (2006), a comunidade de pesquisadores da AI ainda não vê a AI amigável como parte dos problemas. Eu desejaria poder citar uma referência a este efeito, mas, eu não posso citar uma ausência da literatura. A AI amigável está ausente da paisagem conceitual, não apenas impopular ou sem fundamento. Você não pode nem mesmo chamar como “AI amigável” um ponto em branco no mapa, porque

não há nenhuma noção que algo é que faz falta. Se você ler livros populares/semitecnicos que propõe como construir uma AI, tal como Gödel, Escher, Bach (Hofstadter 1979) ou A Sociedade da Mente (Minsky 1986), você pode se recordar que você não viu a AI amigável discutida como parte do desafio. Nem eu vejo a questão sendo discutida na literatura técnica como um problema técnico. Minha tentative de busca de literatura resultou em documentos não técnicos, primeirariamente breves, desconectados entre si, sem nenhuma referência principal em comum exceto as três leis de Isaac Asimov da Robotica". (Asimov 1942.)

Dado que este é o ano de 2006, por que não estão mais pesquisadores de AI falando sobre a segurança? Eu não tenho nenhum acesso privilegiado à psicologia alheia, mas, eu vou especular brevemente baseado em discussões pessoais.

O campo da inteligência artificial adaptou-se a suas experiências nos os últimos cinqüênta anos: em particular, o padrão de grandes promessas, especialmente de potencialidades a nível humano, seguidas pelas embaraçosas falhas públicas. Atribuir este constrangimento à "AI" é talvez injusto; uns pesquisadores mais sábios que não fizeram nenhuma promessa não viram seu conservadorismo fazendo sucesso na mídia. Ainda assim as promessas falhas vêm rapidamente à mente, tanto dentro quanto fora do campo da AI, quando a AI avançada é mencionada. A cultura da pesquisa doeAI adaptou-se a esta circunstância: Há um tabu contra falar sobre potencialidades no nível humanol. Há um tabu mais forte ainda contra a qualquer um que parece ser reivindicando ou predizendo uma potencialidade que não foi demonstrada com um código que rodando. A percepção que eu encontrei é que qualquer um que reivindica pesquisar a AI amigável está reivindicando implicitamente que seu projeto de AI é tão poderoso que precisa ser amigável.

Deveria ser óbvio que isso não é logicamente verdadeiro, nem praticamente uma boa filosofia. Se imaginarmos alguém criando uma AI verdadeira e madura que seja poderosa o suficiente para que precise ser amigável, e, além disso, como é o nosso resultado desejado, esta AI seja realmente amigável, então alguém deve ter trabalhado em AI amigável por anos e anos. AI amigável não é um módulo que você pode inventar instantaneamente no exato momento em que pela primeira vez isso seja necessário, e então anexá-lo em um projeto existente acabado, que de outra forma permaneceria completamente inalterado.

O campo da AI tem técnicas, tais como redes neurais e de programação evolucionária, que cresceram em potencia com a lenta passagem das décadas. Mas as redes neurais são opacas - o usuário não tem idéia de como a rede neural toma as suas decisões - e não pode ser facilmente processada de maneira transparente, as pessoas que inventaram e refinaram as redes neurais não estavam pensando sobre os problemas a longo prazo da AI amigável. A programação evolutiva (PE) é estocástica, e

não preservar com precisão a meta de otimização do código gerado; PE lhe fornece o código que o faz o que você ordenou, na maioria das vezes, nas circunstâncias testadas, mas o código também pode fazer outra coisa em paralelo. EP é uma técnica poderosa, ainda em maturação que é intrinsecamente inadequada para as demandas de AI amigável, que, como já propus, exige ciclos repetidos de recursivo de auto-aperfeiçoamento que preserve com precisão a meta de otimização estável.

As técnicas mais poderosas atuais de AI, uma vez que foram desenvolvidas e, em seguida, refinadas e melhoradas ao longo do tempo, têm incompatibilidades com os requisitos básicos de AI amigável da forma como vejo hoje. O problema Y2K, que foi muito caro para consertar, mas não globalmente catastrófico, analogamente surgiu da falta de previsão de requisitos de design futuros. O cenário de pesadelo é que nós nos encontramos presos a um catálogo de técnicas de AI maduras e poderosas disponíveis publicamente que se combinam para produzir uma AI não amigável, mas que não podem ser usados para construir uma amigável sem refazer as últimas três décadas do trabalho de AI a partir do zero.

No campo da AI é uma ousadia discutir abertamente o nível humano de AI, após as experiências passadas do campo com essa discussão. Há a tentação de congratular-se por ousar tanto, e depois parar. Discutir AI trans-humana parece ridículo e desnecessário depois de tanta ousadia. (Mas não há motivo privilegiado porque as AIs iriam se desenvolver lentamente todo o caminho até a escala de inteligência hmana, e depois param para sempre exatamente nesse ponto). Ousar falar de IA amigável, como precaução contra o risco global catastrófico da AI trans-humana estaria dois níveis acima do nível de ousadia que é apenas ousado o suficiente para ser visto como transgressor e corajoso.

Há também uma objeção pragmática que admite que questão da AI amigável seja um problema importante, mas teme que, dado o nosso estado atual do conhecimento, nós simplesmente não estamos em posição para enfrentá-la: Se nós tentarmos resolver o problema agora, nós apenas falharemos, ou produziremos anticiência, em vez de ciência.

Vale a pena se preocupar com essa objeção. Parece-me que o conhecimento está lá fora, que é possível estudar um conjunto suficientemente grande de conhecimento existente e, em seguida, abordar a AI amigável sem bater de frente em num muro de tijolos, mas o conhecimento está espalhado por várias disciplinas: teoria da decisão, psicologia evolucionária, teoria da probabilidade, biologia evolucionária, psicologia cognitiva, teoria da informação e o campo tradicionalmente conhecido como "Inteligência Artificial”... Não existe currículo que tenha preparado um grande grupo de pesquisadores existentes para fazer avançar uma AI amigável.

A regra de "dez anos" para o gênio, validado através de domínios que vão da matemática à música ao tênis competitivo, afirma que ninguém consegue um desempenho excepcional em qualquer campo sem pelo menos dez anos de esforço. (Hayes 1981).As sinfonias que Mozart começou a compor aos 4 anos, não ‘eram’ as sinfonias de Mozart - que

levaram mais 13 anos para serem compostas. (Weisberg, 1986.) Minha própria experiência com a curva de aprendizagem reforça esta preocupação. Se quisermos que as pessoas façam progresso em AI amigável, então eles têm que começar a treinar, em tempo integral, anos antes de se tornarem urgentemente necessários.

Se amanhã a Fundação Bill e Melinda Gates destinarem cem milhões de dólares do dinheiro de doação para o estudo da AI amigável, em seguida, milhares de cientistas começarão a reescrever as suas propostas de doação para fazê-los relevantes para AI amigável. Mas não estariam genuinamente interessados no problema, testemunho de que eles não mostram curiosidade antes que alguém se ofereça para pagar-lhes. Enquanto uma inteligência artificial geral está fora de moda e a AI amigável totalmente fora do radar, pelo menos podemos assumir que alguém que fale sobre o problema esteja realmente interessado. Se você joga muito dinheiro em um problema que um campo não está preparado para resolver, o excesso de dinheiro é mais suscetível de produzir anti-ciência do que ciência - uma confusão de falsas soluções.

Eu não posso considerar este veredito como uma boa notícia. Nós todos estaríamos muito mais seguros se a AI amigável pudesse ser resolvida acumulando gente e dinheiro. Mas em 2006 eu duvido que este seja o caso, o campo de AI amigável, Inteligência Artificial em si, encontra-se em um estado caótico. No entanto, se se afirma que ainda não podemos fazer progressos sobre AI amigável, sobre o que sabemos muito pouco, devemos perguntar quanto tempo estudaram antes de chegar a esta conclusão. Quem pode dizer o que a ciência não sabe? Há muito tempo, há ciência demais para apenas um ser humano aprender. Quem pode dizer que não estamos prontos para uma revolução científica, um avanço surpresa? E se não pudermos progredir na AI amigável, será porque não estamos preparados, isso não significa que não precisamos dela. Essas duas afirmações não são de todo equivalentes!

Portanto, se acharmos que não podemos fazer progressos na AI amigável, então temos que descobrir como sair deste regime o mais rápido possível! Não há garantia alguma de que, só porque não enfrentamos determinado risco, o risco obedientemente se afaste.

Se cientistas jovens não comprovadamente brilhantes não se interessarem pelo tema AI amigável por suas próprias iniciativas, então, acredito que seria muito benéfico para a espécie humana, se pudéssemos investir em pesquisa para eles estudarem o problema em tempo integral. Alguns fundos para AI amigável são necessários para este resultado, muito mais financiamento que existe atualmente. Mas temo que, nesses estágios iniciais, um Projeto Manhattan só iria aumentar a razão barulho/sinal.

Conclusion

Ocorreu-me certa vez que a civilização moderna ocupa um estado instável. I.J. Good levantou a hipótese de que a explosão de inteligência descreve um sistema dinamicamente instável, como uma caneta precariamente

equilibrada sobre sua ponta. Se a caneta é exatamente vertical, pode permanecer em pé, mas se pender um pouco da vertical a gravidade a puxa nessa direção, e o processo se acelera. Assim, sistemas mais inteligentes com maior facilidade se tornariam ainda mais inteligentes.

Um planeta morto, sem vida orbitando sua estrela, também é estável. Ao contrário de uma explosão de inteligência, a extinção não é um atrator dinâmico - há uma grande diferença entre quase extinto e extinto. Mesmo assim, a extinção total é estável.

Deveria nossa civilização eventualmente vagar para um ou outro modo?

Levando em conta a lógica, o argumento acima contém falhas como na ‘Falácia da Pizza Gigante’, por exemplo: as mentes não vagueiam cegamente por atratores, elas têm motivações. Mesmo assim, eu suspeito que, pragmaticamente falando, nossas alternativas se reduzam a nos tornarmos mais inteligentes ou extintos.

A natureza não é cruel, mas indiferente, uma neutralidade que muitas vezes parece indistinguível de hostilidade. A realidade o joga em um desafio após o outro, e quando você encara um desafio que não pode suportar, você sofre as consequências. Muitas vezes, a Natureza apresenta exigências que são manifestamente abusivas, mesmo em testes onde a pena para o fracasso é a morte. Como é que um camponês medieval do século 10 deve inventar uma cura para a tuberculose? A natureza não apresenta seus desafios conforme a sua habilidade ou seus recursos, ou quanto tempo livre você tem que pensar sobre o problema. E quando lhe for apresentado um desafio letal acima das suas capacidades, você morre. Pode ser desagradável pensar assim, mas essa tem sido a realidade para os seres humanos, por milhares e milhares de anos. A mesma coisa poderia facilmente afetar toda a espécie humana, se a toda a espécie humana enfrentar um desafio injusto.

Se os seres humanos, não envelhecessem, mesmo que os centenários tivessem a mesma taxa de mortalidade dos adolescentes, não seriam imortais. Nós durariamos apenas até as probabilidades nos atingirem. Para viver até um milhão de anos, como um ser humano que não envelhecesse em um mundo tão arriscado como o nosso, você deve de alguma forma reduzir sua probabilidade anual de acidentes a quase zero. Você não pode dirigir você não pode voar você não pode atravessar a rua, mesmo depois de olhar os dois lados, pois ainda é um risco demasiado grande. Mesmo se você abandonou todos os pensamentos de diversão, desistiu de viver para preservar a sua vida, você não pode navegar em uma pista de obstáculos

por milhões de anos. Seria não fisicamente impossível, mas cognitivamente impossível.

A espécie humana, Homo sapiens, não envelhece, mas não é imortal. Os hominídeos sobreviveram por tanto tempo só porque, durante os últimos milhões de anos, não houve arsenais de bombas de hidrogênio, nem naves espaciais para orientar asteróides em direção à Terra, nem laboratórios de armas biológicas para produzir super-vírus, sem perspectiva de retorno anual de guerra nuclear ou uma guerra nanotecnológica, nem uma Inteligência Artificial vagando. Para sobreviver a qualquer momento sensível, precisamos para diminuir cada risco para quase zero. "Bastante bom" não é bom o suficiente para durar mais um milhão de anos.

Parece um desafio injusto. Tal competência não é historicamente típica das instituições humanas, não importa o quanto se esforcem. Durante décadas os E.U. A. e a U.R.S.S. evitaram uma guerra nuclear, mas não perfeitamente, houve ameaças próximas, como a crise dos mísseis cubanos em 1962. Se postularmos que as mentes futuras exibem a mesma mistura de loucura e sabedoria, a mesma mistura de heroísmo e egoísmo, como as mentes que lemos nos livros de história - então o jogo do risco existencial já terminou; perdeu-se desde o início. Podemos sobreviver por mais uma década, até mesmo um século, mas não mais um milhão de anos.

Mas a mente humana não é o limite do possível. O Homo sapiens representa a primeira forma de inteligência geral. Nascemos no princípio das coisas, no amanhecer da mente. Com sorte, os futuros historiadores olharão para trás e descreverão o mundo atual como um mundo estranho, em fase de adolescência, quando a humanidade era inteligente o suficiente para criar enormes problemas para si, mas não inteligente o suficiente para resolvê-los.

No entanto, antes de podermos passar a essa fase da adolescência devemos, como adolescentes, enfrentar um problema de adultos: o desafio da inteligência mais inteligente do que os humanos. Este é o caminho para sair da fase de alta mortalidade do ciclo de vida, o caminho para fechar a janela de vulnerabilidade, é também, provavelmente, o maior risco que já enfrentamos. Inteligência Artificial é uma estrada para este desafio, e acredito que é a estrada que acabaremos tomando. Penso que, no final, provaremos ser mais fácil construir um Boeing 747 a partir do zero, do que ampliar a escala de um pássaro existente ou equipá-lo com motores a jato.

Eu não quero ter a audácia colossal de tentar construir, com um propósito específico, um objetivo definido, algo mais inteligente que nós mesmos. Mas vamos fazer uma pausa e recordar que a inteligência não é a primeira coisa que a ciência humana já encontrou que provou ser difícil de compreender. Estrelas eram mistérios, química e biologia também eram. Gerações de investigadores tentaram e não conseguiu entender os mistérios, que adquiriram a reputação de ser impossível para a ciência. Era uma vez, uma época em que ninguém entendia por que alguma matéria era inerte e sem vida, enquanto outra pulsante de sangue e vitalidade. Ninguém sabia como se reproduzia a matéria viva, ou como nossas mãos obedeciam as nossas ordens mentais. Lord Kelvin escreveu:

"A influência da vida animal ou vegetal na matéria está infinitamente além do alcance de qualquer pesquisa científica até agora desenvolvida. Seu poder de dirigir os movimentos de partículas em movimento, no milagre diariamente demonstrado do nosso livre-arbítrio humano e no crescimento de geração após geração de plantas a partir de uma única semente, são infinitamente diferente de qualquer possível resultado da interação fortuita de átomos. " (Citado em 1912 Macfie).

Toda a ignorância científica é santificada pela antiguidade. Toda e qualquer falta de conhecimento remonta aos primórdios da curiosidade humana e o furo se estende através dos tempos, aparentemente eterno, até alguém preenchê-lo. Eu acho que é possível para simples seres humanos falíveis vencerem o desafio de construir Friendly AI. Mas só se a inteligência deixar de ser um mistério sagrado para nós, como a vida era um mistério sagrado para Lord Kelvin. A inteligência deve deixar de ser qualquer tipo de mistério, seja sagrado ou não. Devemos executar a criação de Inteligência Artificial como a aplicação exata de uma arte exata. E talvez então possamos vencer.

Bibliography

Asimov, I. 1942. Runaround. Astounding Science Fiction, March 1942.

Barrett, J. L. and Keil, F. 1996. Conceptualizing a non-natural entity: Anthropomorphism in God concepts.

Cognitive Psychology, 31: 219-247.

Bostrom, N. 1998. How long before superintelligence? Int. Jour. of Future Studies, 2.

Bostrom, N. 2001. Existential Risks: Analyzing Human Extinction Scenarios. Journal of Evolution and

Technology, 9.

Brown, D.E. 1991. Human universals. New York: McGraw-Hill.

Crochat, P. and Franklin, D. (2000.) Back-Propagation Neural Network Tutorial.

http://ieee.uow.edu.au/~daniel/software/libneural/

Deacon, T. 1997. The symbolic species: The co-evolution of language and the brain. New York: Norton.

Drexler, K. E. 1992. Nanosystems: Molecular Machinery, Manufacturing, and Computation. New York:

Wiley-Interscience.

Ekman, P. and Keltner, D. 1997. Universal facial expressions of emotion: an old controversy and new

findings. In Nonverbal communication: where nature meets culture, eds. U. Segerstrale and P. Molnar.

Mahwah, NJ: Lawrence Erlbaum Associates.

Good, I. J. 1965. Speculations Concerning the First Ultraintelligent Machine. Pp. 31-88 in Advances in

Computers, vol 6, eds. F. L. Alt and M. Rubinoff. New York: Academic Press.

Hayes, J. R. 1981. The complete problem solver. Philadelphia: Franklin Institute Press.

Hibbard, B. 2001. Super-intelligent machines. ACM SIGGRAPH Computer Graphics, 35(1).

Hibbard, B. 2004. Reinforcement learning as a Context for Integrating AI Research. Presented at the 2004

AAAI Fall Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research.

Hofstadter, D. 1979. Gödel, Escher, Bach: An Eternal Golden Braid. New York: Random House

Jaynes, E.T. and Bretthorst, G. L. 2003. Probability Theory: The Logic of Science. Cambridge: Cambridge

University Press.

Jensen, A. R. 1999. The G Factor: the Science of Mental Ability. Psycoloquy, 10(23).

MacFie, R. C. 1912. Heredity, Evolution, and Vitalism: Some of the discoveries of modern research into

these matters – their trend and significance. New York: William Wood and Company.

McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C. E. 1955. A Proposal for the Dartmouth

Summer Research Project on Artificial Intelligence.

Merkle, R. C. 1989. Large scale analysis of neural structure. Xerox PARC Technical Report CSL-89-10.

November, 1989.

Merkle, R. C. and Drexler, K. E. 1996. Helical Logic. Nanotechnology, 7: 325-339.

Minsky, M. L. 1986. The Society of Mind. New York: Simon and Schuster.

Monod, J. L. 1974. On the Molecular Theory of Evolution. New York: Oxford.

Moravec, H. 1988. Mind Children: The Future of Robot and Human Intelligence. Cambridge: Harvard

University Press.

Moravec, H. 1999. Robot: Mere Machine to Transcendent Mind. New York: Oxford University Press.

Raymond, E. S. ed. 2003. DWIM. The on-line hacker Jargon File, version 4.4.7, 29 Dec 2003.

Rhodes, R. 1986. The Making of the Atomic Bomb. New York: Simon & Schuster.

Rice, H. G. 1953. Classes of Recursively Enumerable Sets and Their Decision Problems. Trans. Amer.

Math. Soc., 74: 358-366.

Russell, S. J. and Norvig, P. Artificial Intelligence: A Modern Approach. Pp. 962-964. New Jersey:

Prentice Hall.

Sandberg, A. 1999. The Physics of Information Processing Superobjects: Daily Life Among the Jupiter

Brains. Journal of Evolution and Technology, 5.

Schmidhuber, J. 2003. Goedel machines: self-referential universal problem solvers making provably

optimal self-improvements. In Artificial General Intelligence, eds. B. Goertzel and C. Pennachin. Forthcoming. New York: Springer-Verlag.

Sober, E. 1984. The nature of selection. Cambridge, MA: MIT Press.

Tooby, J. and Cosmides, L. 1992. The psychological foundations of culture. In The adapted mind:

Evolutionary psychology and the generation of culture, eds. J. H. Barkow, L. Cosmides and J. Tooby. New

York: Oxford University Press.

Vinge, V. 1993. The Coming Technological Singularity. Presented at the VISION-21 Symposium,

sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute. March, 1993.

Wachowski, A. and Wachowski, L. 1999. The Matrix, USA, Warner Bros, 135 min.

Weisburg, R. 1986. Creativity, genius and other myths. New York: W.H Freeman.

Williams, G. C. 1966. Adaptation and Natural Selection: A critique of some current evolutionary thought.

Princeton, NJ: Princeton University Press.